Une Méthodologie de Recommandations Produits Fondée sur l'Actionnabilité et l'Intérêt Économique des Clients - Application à la Gestion de la Relation Client du groupe VM Matériaux

Thomas Piton

UNIVERSITÉ DE NANTES

École polytechnique de l'Université de Nantes

ÉCOLE DOCTORALE

« SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET MATHÉMATIQUES »

Année : 2011 N° attribué par la bibliothèque

Une Méthodologie de Recommandations Produits Fondée sur l'Actionnabilité et l'Intérêt Économique des Clients

Application à la Gestion de la Relation Client du groupe VM Matériaux

---

THÈSE DE DOCTORAT

Discipline : FOUILLE DE DONNÉES
Spécialité : INFORMATIQUE

Présentée
et soutenue publiquement par

Thomas PITON

Le 13 Octobre 2011, devant le jury ci-dessous

Président	:	Djamel Abdelkader ZIGHED	Professeur, Université Lumière Lyon 2
Rapporteurs	:	Gilbert SAPORTA	Professeur, CNAM de Paris
		Jérôme DARMONT	Professeur, Université Lumière Lyon 2
Examinateurs	:	Julien BLANCHARD	Maître de Conférences, Polytech'Nantes
		Henri BRIAND	Professeur Émérite, Polytech'Nantes
Invités	:	Pierrick RICHARD	DSI, groupe VM Matériaux
		Gaëtan BLAIN	Adjoint DSI, groupe VM Matériaux

Directeur de thèse : **Fabrice GUILLET**

ED : 503-142

Typeset in LaTeX 2ε.

À Marion.

Remerciements

Ces trois années de thèse représentent pour moi une expérience enrichissante sur le plan scientifique avec des échanges intéressants et constructifs et des discussions riches avec différents collègues de travail.

Un début de carrière dans le monde industriel, celui du négoce de matériaux, passionnant, motivant et rebondissant avec ses évolutions et ses imprévus. Les gens compétents et chaleureux que j'ai côtoyés, m'ont enrichi humainement et professionnellement. L'aide et le soutien quotidien de ces personnes ont amélioré mon savoir-faire, mon savoir-être et mes connaissances.

Mes remerciements et ma gratitude vont à mon directeur de thèse Fabrice Guillet. Il a su me donner les bonnes directives et me faire profiter de son expérience de la recherche tout en m'accordant une grande liberté de travail. Je remercie mon encadrant Julien Blanchard pour ses capacités pédagogiques qui m'ont permis de développer mes aptitudes pour la recherche, je lui en suis très reconnaissant. Je remercie également Henri Briand pour sa vision avant-gardiste et fondatrice de ma thèse et du partenariat entre le laboratoire LINA et le groupe VM Matériaux. Je remercie Gilbert Saporta et Jérôme Darmont d'avoir accepté d'être les rapporteurs de ma thèse, et pour l'attention avec laquelle ils l'ont lue et évaluée. Je remercie Djamel Abdelkader Zighed d'avoir accepté de faire partie de mon jury.

Merci aussi à mon directeur informatique Pierrick Richard pour son pragmatisme et sa vision à la fois professionnelle et scientifique de l'informatique. Je tiens à remercier Gaëtan Blain pour ses grandes qualités méthodologiques et ses conseils dans les moments difficiles. Je tiens également à remercier le directoire du groupe VM Matériaux pour la confiance qu'ils ont su m'apporter tout au long de ces trois années.

Je remercie Pascale Kuntz de m'avoir accueilli au sein de l'équipe COD ainsi que Claudia Marinica et toute l'équipe COD. Je remercie également très chaleureusement l'ensemble du service informatique de VM Matériaux et plus particulièrement mes deux collègues Sylvain Richard et Emmanuel Richard qui ont su m'épauler et m'orienter dans de nombreux projets. Je remercie Catherine Galais pour ses relectures minutieuses des articles anglophones. Je remercie Laurent Tessier pour la passion qu'il a su me transmettre. Enfin, je remercie toute ma famille, mes parents et plus particulièrement mon amie Marion pour son dévouement, sa patience, et sa confiance tout au long de ces trois années.

Table des matières

Liste des figures

Liste des tableaux

1 Introduction

SOMMAIRE

1.1 Contexte

Dans un contexte économique de plus en plus concurrentiel, la richesse des entreprises réside dans leurs clients. La part de clients a remplacé la part de marché. La plupart des entreprises savent qu'il existe d'importantes différences de rentabilité entre les clients existants et les nouveaux clients. En effet, il est moins coûteux de vendre à un client existant qu'à de nouveaux clients car les entreprises n'ont pas à prouver leur notoriété, ni à supporter les coûts des programmes de démarchage de nouveaux clients. Une étude comparant des échantillons comptant jusqu'à 50 000 clients, a révélé que le retour sur investissement de produits marketing spécifiques était de 530 % auprès de clients existants contre -30 % auprès de prospects [150]. Tout ceci explique pourquoi les entreprises cherchent à fidéliser leurs clients existants en entretenant des relations plus étroites. La mise en œuvre d'une stratégie de fidélisation se base principalement sur la relation avec le client. La Gestion de la Relation Client ou *Customer Relationship Management* (CRM) [42, 157] utilise stratégiquement l'information, les processus, la technologie et les personnes pour gérer la relation entre le client et l'entreprise. Plus particulièrement le CRM analytique analyse des données clients et achats issues des applications opérationnelles et offre une vision unifiée du comportement des clients.

Cette approche de la connaissance client fondée sur l'analyse des données se heurte toutefois au volume des données qui croît de façon très importante. Par exemple, *Yahoo !* produit environ seize milliards de transactions par jour, représentant environ dix téraoctets de données [88]. Un grand détaillant de plus de 3 000 magasins génère 300 millions d'événements chaque jour à partir de puces RFID[1] [115]. Le réseau social de messagerie instantanée de *Microsoft* dispose de plus de 250 millions de nœuds [149]. Face à cette abondance d'informations stockées dans les bases de données [151], les entreprises ont mis en œuvre des outils et des techniques pour améliorer leur connaissance des clients : l'ECD ou l'Extraction de Connaissances à partir des Données. Elles entendent utiliser les grandes quantités d'information qu'elles possèdent et les tourner en avantages compétitifs. L'ECD est un processus non trivial d'identification de connaissances nouvelles, valides et potentiellement utiles pour les experts métier [90, 95]. Ce domaine d'étude emprunte à la fois à la statistique, à l'analyse de données et à l'intelligence artificielle [116, 118]. La fouille de données [179] est une étape moteur de l'ECD trouvant de nombreuses applications dans l'industrie et le commerce [155], notamment pour le CRM [235, 275]. Les experts métier utilisent la connaissance extraite des modèles d'ECD pour comprendre le comportement des clients et prendre des décisions à bon escient (paradigme de l'*actionable knowledge* [52, 104]). L'actionnabilité de la connaissance [14, 82, 252] demeure aujourd'hui un verrou scientifique en ECD, elle est cependant cruciale pour l'usage des modèles. Elle nécessite d'être complétée par la mesure du profit [83, 251].

Dans ce double contexte, nous nous intéressons ici aux systèmes de recommandation. Les systèmes de recommandation [237] permettent d'offrir de manière proactive des suggestions personnalisées et adaptées aux besoins du client en fonction de

1. *Radio Frequency IDentification* ou Identification par Radio Fréquence.

son comportement, à l'aide par exemple de ventes croisées (*cross-selling*) ou de montées en gammes (*up-selling*). Recommander des produits et des services peut renforcer la relation entre l'acheteur et le vendeur et donc augmenter les bénéfices [309]. La croissance explosive d'Internet et l'émergence du commerce électronique ont conduit à l'essor des systèmes de recommandation [236], notamment dans le domaine du CRM [156, 249]. Ces dernières années sont révélatrices de l'utilisation des systèmes de recommandation sur Internet à travers les films [2], les livres [3] et la musique [4]. Les systèmes de recommandation s'appuient couramment sur des méthodes d'apprentissage ou de fouille de données [238].
Dans le contexte du commerce électronique, le fournisseur cherche en permanence à offrir une expérience interactive plus large à l'ensemble des internautes. Ceci est permis par le coût de production très faible d'une recommandation et par l'impact insignifiant d'une mauvaise recommandation. *Amazon* a même reconnu avoir eu recours à de « fausses » recommandations afin de favoriser la sortie de nouveaux articles valorisant ses partenariats fournisseurs [297]. Les recommandations sont qualifiées de non intrusives et ne nécessitent ni d'explication pour le client, ni d'argumentaire pour le vendeur. Ces systèmes réalisent majoritairement des recommandations fondées sur le produit acheté par le client au moment présent [2]. Ce modèle convient parfaitement aux sites de commerce en ligne. En revanche, l'adaptation des systèmes de recommandation à des processus de vente assistés par un commercial constitue de notre point de vue un second verrou scientifique.

Dans le cadre d'une stratégie commerciale basée sur les forces de vente, comment fidéliser les clients pour accroître leur valeur ?

Les travaux présentés ici s'inscrivent dans le domaine de la gestion de la relation client pour une stratégie commerciale de fidélisation. Nous proposons d'adopter une philosophie différente des systèmes de recommandation existants : dans un contexte de démarchage par un commercial, quel(s) produit(s) recommander au client pour déclencher l'acte d'achat ? Comment identifier les manques à gagner, souvent synonyme d'achats à la concurrence ? Effectuer ce type de recommandation dite « intrusive » nécessite (i) d'identifier des comportements typiques d'achat sur le long terme, et (ii) de les appliquer au client en veillant à respecter son comportement. Autrement dit, il s'agit d'identifier les couples *(client, produit/service)* présentant un potentiel de marge de progression de chiffre d'affaires. Le coût d'une mauvaise recommandation s'avère plus élevé dans le cadre d'une visite commerciale que dans celui d'un site de commerce en ligne [169]. Le commercial peut refuser d'utiliser le système s'il ne juge pas les recommandations suffisamment pertinentes. Le client peut ne pas apprécier la recommandation et les impacts sur la fidélisation peuvent être dramatiques. Notre démarche s'inscrit dans un contexte de recommandation dite « intrusive » et une politique commerciale qualifiée d'« agressive », proposant au commercial des recommandations actionnables et explicables.

2. http://www.netflix.com
3. http://www.amazon.com
4. http://www.last.fm

1.2 Contribution de la thèse

Notre problématique se décline en particulier dans le contexte industriel, où le défi majeur des systèmes de recommandation reposant sur la fouille de données est de passer de l'étude des algorithmes à la formalisation d'un savoir actionnable et profitable pour les experts métier [104]. La connaissance extraite des modèles de fouille de données doit permettre aux experts de prendre des décisions. Pour rationaliser la prise de décision, nous intégrons son impact financier. Nous proposons une méthodologie pour les systèmes de recommandation fondée sur l'analyse des chiffres d'affaires des clients à différents niveaux d'une taxonomie produits. Plus précisément, la méthodologie consiste à extraire des comportements de référence sous la forme de règles et à en évaluer l'actionnabilité et l'intérêt économique. Les recommandations sont réalisées en ciblant les contre-exemples actionnables sur les règles les plus rentables. Cette approche a été mise en œuvre au sein de l'outil *ARKIS* (*Association Rule Knowledge Interactive Search*) développé dans le cadre de cette thèse. Les contributions principales de la thèse sont résumées comme suit :

- Nous développons une nouvelle **méthodologie** pour la recommandation de produits à partir de chiffres d'affaires ;
- Nous proposons une mesure de l'**actionnabilité** des recommandations fondée sur la similarité entre les clients exemples et les clients contre-exemples ;
- Nous offrons une mesure originale de l'**intérêt économique** des règles pour la recommandation, fondée sur des critères définis par les experts métier ;
- Nous mesurons l'**efficacité** de notre méthodologie sur le jeu de données de référence *MovieLens* à l'aide d'une validation croisée ;
- Nous appliquons notre méthodologie sur une **base de données réelle** du groupe VM Matériaux composée de plus de 10 000 clients et 100 000 produits.

1.3 Organisation du document

Cette thèse est organisée comme suit :
Le **deuxième chapitre** est consacré à l'*actionnabilité* et plus précisément à l'extraction de connaissances actionnables et profitables pour les experts métier. La technique d'extraction de règles d'association actionnables est approfondie.

Le **troisième chapitre** est dédié à la présentation des systèmes de recommandation. Nous nous concentrons davantage sur les techniques répondant à l'approche de filtrage collaboratif pour l'extraction de recommandations dites personnalisées.

Le **quatrième chapitre** développe notre méthodologie *CAPRE* (*Customer Actionability and Profitability Recommendation*) d'extraction de recommandations actionnables et profitables à partir des chiffres d'affaires de produits.

Le **cinquième chapitre** implémente notre méthodologie au travers de l'outil *ARKIS*. Nous présentons les fonctionnalités puis leur implémentation et exposons quelques exemples d'utilisation d'*ARKIS*. Nous validons et discutons notre méthodologie au travers d'une validation croisée sur le jeu de données de référence *MovieLens*.

Le **sixième chapitre** applique notre méthodologie aux données réelles du groupe VM Matériaux. Une phase préparatoire illustre la mise en place d'un modèle d'appétence clients pour une opération commerciale. Nous appliquons et validons ensuite notre méthodologie *CAPRE* sur les clients ciblés, fournissant ainsi aux commerciaux des recommandations explicites sur les clients présentant un manque à gagner.

1.4 Liste des publications

Conférences Internationales

- [225] T.PITON, J.BLANCHARD et F.GUILLET. **CAPRE : A New Methodology for Product Recommendation Based on Customer Actionability and Profitability**. *In proceedings of the fifth International Workshop on Domain Driven Data Mining (DDDM 2011) in conjunction with IEEE ICDM 2011, December, 2011, Vancouver, Canada, IEEE Computer Society*.

- [223] T.PITON, J.BLANCHARD, H.BRIAND et F.GUILLET. **Domain Driven Data Mining to Improve Promotional Campaign ROI and Select Marketing Channels**. *In The 18th ACM Conference on Information and Knowledge Management, pages 1057-1066, Hong-Kong, 2009*.

Conférences Nationales

- [226] T.PITON, J.BLANCHARD et F.GUILLET. **Une Méthodologie de Recommandations Produits Fondée sur l'Actionnabilité et l'Intérêt Économique des Clients**. *Actes des onzièmes journées Extraction et Gestion des Connaissances EGC'2011, volume E-20 RNTI, pages 203-214, Brest France, 2011*.

- [224] T.PITON, J.BLANCHARD, H.BRIAND, L.TESSIER et G.BLAIN. **Analyse et Application de Modèles de Régression Pour Optimiser le Retour sur Investissement d'Opérations Commerciales**. *Actes des neuvièmes journées Extraction et Gestion des Connaissances EGC'2009, RNTI, pages 25-30, Strasbourg France, 2009*.

Prix

- T.PITON. **Extraction de Connaissances Actionnables à l'Aide de Techniques de Fouille de Données pour Améliorer la Fiabilité des Connaissances et Optimiser les Performances**. *Journée des doctorants de l'école doctorale STIM (JDOC'10), prix de la meilleure présentation orale, Nantes France, Avril 2010*[5].

5. http://edstim.univ-nantes.fr/jdoc2010/index.htm

2

Actionnabilité des connaissances

Business people are not informed about how and what to do to take over the technical deliverables [...]

Longbing Cao, 2008

SOMMAIRE

2.1 Introduction : « découvrir pour agir économiquement »

2.1.1 Information, connaissance et actionnabilité

Une **information** est fondamentalement une action en devenir pour qui sait la mettre en perspective, elle procure la capacité à mettre en œuvre des actions en vue d'influer sur l'environnement [56]. Néanmoins l'information est périssable, sa valeur diminue avec le temps. Plus grand est le nombre d'informations révélées, plus grande est l'intelligence du domaine, jusqu'à ce qu'un certain niveau de compréhension soit atteint : **la connaissance**. La connaissance est donc « le résultat d'un assemblage d'informations traitées auquel l'esprit humain a pu assigner un sens » [180]. Connaissance et information sont deux notions souvent confondues en ECD, l'information n'étant qu'un élément de connaissance susceptible d'être codé, mémorisé et traité [151].
Le néologisme **connaissance actionnable** ou *actionable knowledge* a été introduit dans la littérature par Schön en 1983, afin de dépasser le distinguo habituel entre savoir et savoir-faire (cf. figure 2.1). Dès 1974, Argyris propose avec Schön des modèles organisationnels pour l'action et introduit pour la première fois en 1995 l'expression « savoir pour agir » (*knowledge for action*) [12]. Il ressort de ces travaux que la réflexion sur ses actions, sur ses savoirs d'expérience, entraîne le praticien (e.g. l'expert métier) a mieux prendre conscience des stratégies d'actions qu'il a élaborées, et donc qu'il pourra améliorer. Lorsque l'interprétation d'une information tend à la rendre utile à l'action, on parle de *connaissance actionnable* [12]. La notion de connaissance ou savoir actionnable a été définie par Argyris en 1993 [11] comme « un savoir à la fois valable et pouvant être mis en action dans la vie quotidienne ».

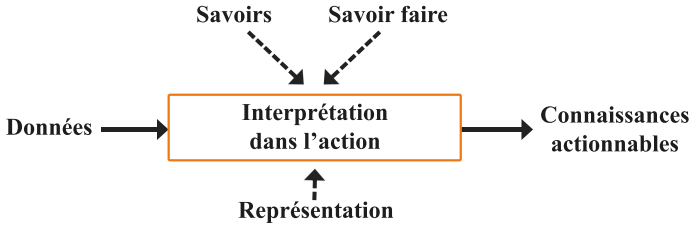


FIGURE 2.1 : Des données aux connaissances actionnables [12]

La gestion des connaissances actionnables est définie comme un « processus de partage dynamique de connaissances utiles à l'action » [11]. Généralement, les outils et les techniques de fouille de données se concentrent principalement sur la découverte de modèles satisfaisant des indicateurs, mais non les experts métier. Ces derniers ont souvent de nombreuses difficultés pour juger ce qui est intéressant et actionnable dans le cadre imparti [289]. Ces constats mettent en évidence qu'il existe un écart important [52, 56, 60, 61] entre les domaines académique et industriel, ainsi qu'entre les *data miners* et les experts métier [220]. Il est essentiel d'élaborer des méthodologies et des techniques efficaces pour combler cet écart et rendre actionnable [59], opérable ou exploitable la connaissance issue de la fouille de données pour les experts métier : c'est l'AKD ou *Actionable Knowledge Discovery* [63].

2.1.2 La valeur stratégique des connaissances

Face au contexte économique actuel, de nombreuses entreprises sont amenées à réexaminer leur stratégie. Bon nombre d'entre elles cherchent à recentrer leurs services sur les besoins de leurs clients et à établir des relations personnalisées [286]. En effet, il est avantageux de connaître ses clients et leurs habitudes d'achat, tout comme il est indispensable de les fidéliser. La plupart des entreprises savent qu'il existe d'importantes différences en termes de chiffre d'affaires et de rentabilité entre les clients fidèles et le reste de la clientèle (souvent appelé les « ventres-mou »). La loi de *Pareto* [160] illustre que généralement, 20 % des clients représentent environ 80 % de l'activité de l'entreprise. Une étude portant sur les clients d'une entreprise de ventes au détail a révélé que les dix premiers pour cent des clients représentaient la moitié du chiffre d'affaires [165]. Une enquête du même type, menée pour les entreprises spécialisées en matériel et logiciels informatiques, a relevé que les cinq premiers pour cent des clients généraient des marges 10 à 50 % plus élevées [286]. De la même manière, un grossiste a découvert qu'environ un tiers de ses clients étaient peu rentables [206].
Les programmes de fidélisation des clients portent généralement leurs fruits. En effet, les données publiques d'*Amazon*, *Ameritrade*, *eTrade*, *Capitol One* et *eBay* ont révélé qu'1 % d'amélioration de la fidélisation des clients générait une augmentation de 5 % des valeurs de la société [111]. Une étude menée dans le secteur des services a montré qu'une amélioration de 5 % de la fidélisation des clients dopait les profits de 25 à 85 % [233]. Une société de B2B (*Business to Business*) a étudié les impacts d'une baisse de 25 % des dépenses consacrées au démarchage et à la fidélisation sur une population de 12 000 clients. Alors que la baisse du retour sur investissement marketing n'était que de 3 % pour le démarchage de nouveaux clients, le taux de fidélisation a chuté quant à lui de 55 % [234].
De nombreuses études ont démontré les impacts de la fidélisation des client existants sur les résultats des entreprises. Une enquête menée auprès de plus de 100 entreprises de divers secteurs d'activité a montré qu'une baisse de la fidélisation client a un impact très élevé sur le développement commercial des entreprises [8]. Une étude du cabinet *McKinsey & Company* [1] a révélé que 70 % des clients qui se tournent vers la concurrence le font à cause de la médiocrité du service. Un rapport *Capgemini* sur le secteur de l'assurance a souligné que 40 % des souscripteurs changeaient d'assureur en cas de mauvaises relations entre les deux partis [283].

2.1.3 Optimisation de la gestion de la relation client

Les entreprises des secteurs qui connaissent le changement le plus rapide, tels que les secteurs de la banque et de la finance, de l'automobile et de la santé, ainsi que de la grande distribution passent d'un modèle axé sur les produits à un modèle axé sur les clients. Pour ces entreprises, connaître le client est devenu la priorité. En entretenant des relations plus étroites avec les clients, elles s'attachent à les fidéliser, ce qui leur permet ensuite de multiplier les ventes et d'accroître les opportunités de

1. http://www.mckinsey.com

ventes croisées ou de montées en gammes. Par exemple, la grande distribution a besoin de connaître ses clients en (i) créant des relations privilégiées sur le modèle du *commerce de quartier* et en (ii) évaluant ses clients dans la durée [275]. Dans un petit commerce de proximité, le vendeur observe le comportement de ses clients et se souvient aisément de ses achats et préférences. A contrario, dans une entreprise de taille importante, le client rentre rarement en contact avec le même employé. Cependant, il est possible d'exploiter les nombreuses transactions laissées par le client.
Le cabinet *Micropole Univers* a réalisé une étude [190] présentant les neuf principaux enjeux commerciaux (cf. figure 2.2) de la gestion de la relation client en entreprise. La connaissance du client est devenue la principale motivation. La figure 2.2 souligne que l'enjeu « Meilleure connaissance client » est sélectionné par 90 % des répondants dont 40 % en rang 1. Les enjeux N° 5 (*Base de données clients commune*) et N° 8 (*Pérennité de la base client*) renforcent ce constat.

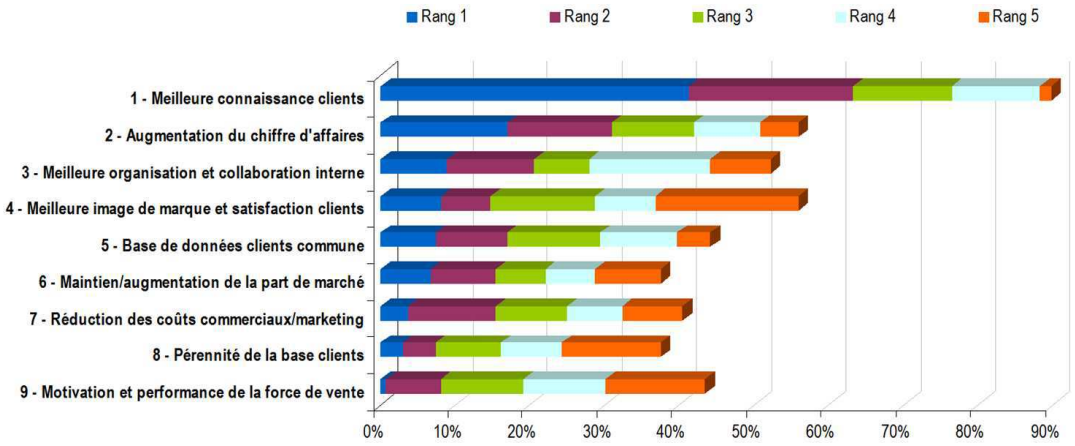


FIGURE 2.2 : La connaissance client : principale motivation des projets CRM [190]

Dans ce contexte, la compréhension des attentes des clients et l'anticipation de leurs besoins deviennent un objectif majeur de nombreuses entreprises, qui souhaitent augmenter la rentabilité [170] et la fidélité de leurs clients [112], tout en maîtrisant les risques et en utilisant les bons canaux de communication au moment opportun. C'est à cet objectif que souhaite répondre le CRM [42, 213]. Bien que le CRM soit maintenant largement reconnu comme une importante approche commerciale, il n'existe pas de définition universellement acceptée du CRM [172, 204]. Voici les définitions les plus couramment rencontrées dans la littérature :

- R.Swift [267], définit page 12 de son ouvrage le CRM comme « une approche d'entreprise à comprendre et à influencer le comportement des clients au moyen de communications utiles en vue d'améliorer l'acquisition, la fidélisation, et la rentabilité des clients » ;

- J.W.Kincaid [148] définit page 43 de son ouvrage le CRM comme « l'utilisation stratégique de l'information, des processus, de la technologie, et des personnes

pour gérer la relation du client avec l'entreprise dans l'ensemble du cycle de vie du client » ;

- A.Parvatiyar et N.Sheth [213] définissent page 5 de leur ouvrage le CRM comme « une stratégie globale pour l'acquisition, le maintien et les partenariats avec les clients ciblés créant de la valeur supérieure pour l'entreprise ».

Le CRM se décompose en deux principaux éléments : le CRM opérationnel et le CRM analytique [28, 120, 269] (cf. figure 2.3) (considérant que le CRM collaboratif est inclus dans le CRM opérationnel).

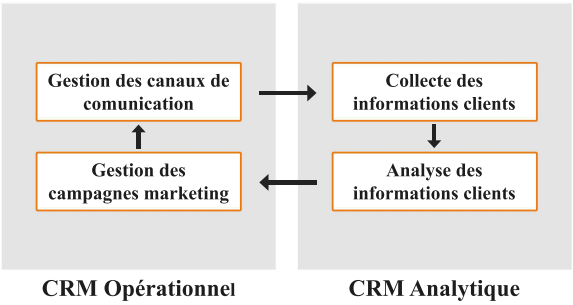


FIGURE 2.3 : La chaîne de la relation client [42]

L'objet du CRM analytique est l'extraction, le stockage, l'analyse et la restitution des informations pertinentes permettant de fournir une vision complète et unifiée du client dans l'entreprise. La matière première du CRM analytique est la donnée, et ses composants sont le *datawarehouse*, le *datamart*, l'analyse multidimensionnelle (*OLAP*[2]), la fouille de données et les outils de *reporting* [29].
Les outils de front-office, la gestion de la force de vente, le service client et le centre d'appel, la gestion des différents canaux de communication (forces de vente, Internet, téléphone, emails, etc.) [45, 79] et la gestion des campagnes marketing pour une mise en œuvre optimale des stratégies identifiées grâce au CRM analytique relèvent du CRM opérationnel. Le CRM opérationnel, qui s'appuie sur les résultats du CRM analytique [126], alimente en retour le CRM analytique en données à analyser. Il y a donc une noria de données entre les CRM (cf. figure 2.3), d'autant plus que la multiplication des canaux de communication fait que l'information client est de plus en plus riche et complexe à collecter et à analyser [275].
Actuellement, de nombreuses entreprises collectent et stockent une multitude de données sur leurs clients, prospects, fournisseurs et partenaires commerciaux. Toutefois, l'incapacité à découvrir de précieuses informations cachées dans les données empêche les organisations de transformer ces données en valeur ajoutée, c'est-à-dire en connaissances actionnables [28].

2. *OnLine Analytical Processing*, un cube à n dimensions (« hypercubes ») dont tous les points dans l'espace multidimensionnel sont précalculés de façon à fournir une réponse très rapide à des questions portant sur plusieurs axes, tel que le chiffre d'affaires par type de client et ligne de produit.

2.1.4 Extraction de connaissances à partir des données

Enjeux

L'Extraction de Connaissances à partir des Données (ECD) se définit comme « l'acquisition de connaissances nouvelles, intelligibles et potentiellement utiles à partir de faits cachés au sein de grandes quantités de données » [90]. On cherche principalement à isoler des traits structuraux (*patterns*) qui soient valides, non triviaux, nouveaux, utilisables et si possible compréhensibles ou explicables. L'ECD impacte de nombreux domaines [275], qui vont de l'infiniment petit (génomique) à l'infiniment grand (astrophysique), du plus ordinaire (gestion de la relation client) au moins habituel (aide au pilotage aéronautique), du plus ouvert (e-commerce) au plus sécuritaire (prévention du terrorisme), du plus industriel (pilotage de la production) au plus académique (enquêtes en sciences humaines), et du plus alimentaire (études agroalimentaires) au plus divertissant (prédiction d'audience à la télévision). La découverte de connaissances devient un enjeu stratégique pour les entreprises [71]. Un large panel de définitions de l'ECD est disponible dans la littérature :

- S.Tufféry [275] : « l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de grandes bases de données, de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données permettant d'étayer des prises de décisions » ;
- U.Fayyad [95] : « l'extraction d'informations originales, auparavant inconnues, potentiellement utiles à partir des données » ;
- K.Parsaye [212] : « un processus d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données » ;
- M.Jambu [140] : « un processus non élémentaire de mises à jour de relations, corrélations, dépendances, associations, modèles, structures, tendances, classes, facteurs obtenus en naviguant à travers de grands ensembles de données ».

Processus

Le processus d'extraction de connaissances (cf. figure 2.4) est généralement découpé en cinq phases [36, 90] :

1. **Sélection** : recensement des données utiles, accessibles (internes ou externes à l'entreprise), légalement et techniquement exploitables, fiables et suffisamment à jour ;
2. **Pré-traitement** : exploration et mise en forme des données pour fiabiliser, remplacer ou supprimer les données incorrectes, soit qu'elles aient trop de valeurs manquantes, des valeurs aberrantes, ou qu'elles aient des valeurs extrêmes (*outliers*) s'écartant trop des valeurs habituellement admises ;

3. **Transformation** : création d'indicateurs pertinents à partir des données brutes et le cas échéant corrigées. Elle peut se faire en remplaçant des grandeurs absolues par des ratios, en calculant des évolutions temporelles de variables, en effectuant des combinaisons linéaires de variables, en composant des variables avec d'autres fonctions, en recodant certaines variables, etc. ;

4. **Élaboration du modèle** : cette étape constitue le cœur de l'activité de la fouille de données. Elle peut constituer un calcul de score, ou plus généralement d'un modèle prédictif. Au cours de cette étape, plusieurs modèles sont généralement élaborés, le plus performant étant choisi ;

5. **Interprétation et évaluation** : on réitère les étapes précédentes jusqu'à obtention de résultats complètement satisfaisants pour les experts métier. Le modèle choisi sera interprété puis validé par les experts.

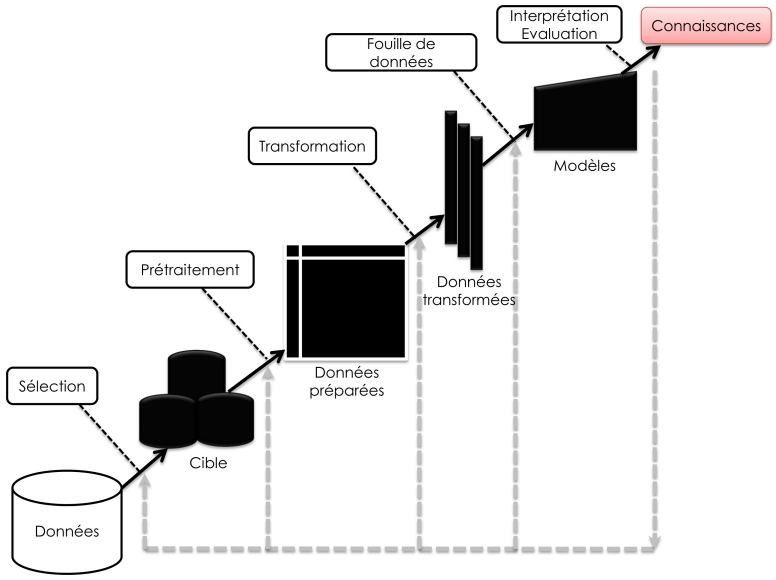


FIGURE 2.4 : Processus d'Extraction de Connaissances à partir des Données [90]

La fouille de données [118, 179] est une étape moteur de l'ECD consistant à identifier les motifs structurant les données. Cette étape est soit descriptive, soit prédictive [70, 178] : les techniques descriptives (ou exploratoires) visent à mettre en évidence des informations présentes mais enfouies sous le volume des données ; les techniques prédictives (ou explicatives) visent à extrapoler de nouvelles informations à partir des informations présentes, ces nouvelles informations pouvant être qualitatives (classement ou *scoring*) ou quantitatives (prédiction). Silberschatz et Tuzhilin comparent en 1995 trois catégories de techniques de découverte de connaissances, plaçant la fouille au cœur de chaque processus.
Cette approche se heurte toutefois à la quantité de données qui croît de manière exponentielle. Le volume des données représente un coût non négligeable pour les entreprises qui sont obligées de plus en plus d'industrialiser leurs modèles pour

répondre aux demandes métier [94]. Du fait de la grande quantité de règles que produisent les algorithmes de fouille de données, le post-traitement est une étape nécessaire.
Différentes solutions ont été proposées pour accompagner les experts métier dans la validation des connaissances ou guider les utilisateurs finaux dans la navigation vers les connaissances. Actuellement, l'essor de l'informatique décisionnelle au sein des entreprises, notamment des petites et moyennes entreprises [103], a conduit à l'expansion de la présentation des connaissances sous forme de requêtes *ad-hoc*, de rapports ou de cubes multidimensionnels pour impliquer au maximum les experts métier [76, 91]. De nombreuses avancées ont été réalisées autour de l'intégration des connaissances dans les cubes multidimensionnels [26, 187] et la restitution sous forme de tableaux de bords [92, 143]. Enfin, la fouille de données est généralement complétée par des méthodes de post-traitement telles que la visualisation [89, 144] et l'utilisation des mesures d'intérêt [106, 109]. L'informatique décisionnelle (ou SIAD [3]) est au cœur de l'évolution allant d'un traitement classique des caractères *a priori* sans signification, vers une manipulation des connaissances. L'entrepôt de données ou *datawarehouse* constitue la pierre angulaire du SIAD.

Applications

Les méthodes de fouille de données sont utilisées dans des applications multiples et dans des domaines très variés [58, 217, 295] :

- Le **secteur bancaire** avec de nombreuses méthodes de *scoring* dues aux actualités bancaires du moment : par exemple, la réforme du ratio de solvabilité qui a donné un grand essor au développement des modèles de risque [275, 276] ;
- La **grande distribution** avec le développement des cartes privatives leurs permettant de se constituer de grandes bases de données enrichies par les informations comportementales provenant des tickets de caisses [27, 171, 291] ;
- La **téléphonie fixe** avec l'ouverture à la concurrence du marché européen et l'arrivée à maturité puis à saturation du marché de la téléphonie mobile, ont avivé des problèmes de départ de clients à la concurrence (*churning*) ;
- L'**industrie automobile** avec le score de ré-achat d'un véhicule de la marque ou la recherche de facteurs à l'origine des défauts de fabrication [133, 155] ;
- La **Vente Par Correspondance** (VPC) afin d'optimiser les ciblages et d'en réduire les coûts, qui peuvent être énormes lorsqu'un catalogue de mille pages en couleur est adressé à plusieurs dizaines de millions de clients. Si la banque est à l'origine du score de risque, la VPC fait partie des premiers secteurs à avoir eu recours au score d'appétence [171] ;
- Le **secteur médical** tant dans les applications descriptives que prédictives. Par exemple, la détermination de groupes de patients susceptibles d'être soumis

3. Système Informatique d'Aide à la Décision

à des protocoles thérapeutiques déterminés, des associations de médicaments en vue notamment de détecter des anomalies de prescription, la recherche de facteurs (marqueurs) susceptibles d'expliquer certaines pathologies [200] ;

- L'**industrie agroalimentaire** ou la bioinformatique avec les analyses sensorielles qui croisent les données sensorielles perçues par les consommateurs avec les mesures instrumentales [22] ;
- Le **Web** avec le commerce électronique [152] et ses nombreuses problématiques émergentes de ventes croisées (*cross-selling*) et de montées en gamme (*up-selling*), de promotions et de recommandations de plus en plus personnalisées pour l'internaute [263, 303].

Acteurs

Le projet de fouille de données doit être supporté par une volonté d'entreprise. Les décideurs ou experts métier doivent être sensibilisés et fortement impliqués pour soutenir le projet [275] tout au long des cinq étapes citées dans la partie 2.1.4.

- En amont, pour définir la problématique, signaler les sources de données intéressantes, fournir les informations nécessaires à l'étude ;
- Pendant l'étude, pour juger de la pertinence et des points à approfondir où à affiner avec le statisticien ;
- En aval, pour utiliser les résultats et mettre en œuvre les actions appropriées.

Il faut associer spécialistes métier et futurs utilisateurs au déroulement de l'étude : les connaissances dans les données ne sont qu'une partie du savoir de l'entreprise, la fouille de données à elle seule ne fournira pas les meilleurs modèles ; ceux-ci naissent du rapprochement entre les connaissances extraites des données et celles tirées de l'expérience des experts métier [275]. Pour que la communication puisse s'appuyer sur les modèles de fouille de données, une telle approche nécessite d'accorder un soin tout particulier à la qualité des modèles produits, par exemple par des méthodes d'évaluation intelligibles [39, 109] et des techniques de représentation telles que celles présentées dans les travaux de Card et al. [64], de Simoff [261] et de Keim et al. [145].

2.1.5 CRM : secteur applicatif de fouille en pleine expansion

Durant ces dernières années, la fouille de données a trouvé de nombreuses applications dans l'industrie et le commerce [155, 171], notamment pour le CRM [235], créant, développant et entretenant des relations personnalisées avec les clients [241]. En outre, la croissance rapide des nouvelles technologies de l'information a considérablement accru les possibilités pour le marketing et a transformé les relations existantes entre les entreprises et leurs clients [204]. Un sondage [4] effectué en 2009 puis

4. http://www.kdnuggets.com/polls/2010/analytics-data-mining-industries-applications.html

en 2010 sur plus de 200 entreprises révèle les principaux secteurs utilisant la fouille de données. Le *CRM* et le *marketing direct* figurent parmi les domaines d'application les plus importants.
Toutefois, la fouille de données pour le CRM a subi un succès limité [112]. Dans les grandes entreprises, il est difficile de collecter et transformer des données [229]. Il existe souvent une limite forte du nombre de clients que l'entreprise peut prendre en compte. De plus, pour chaque client, un grand nombre d'actions possibles peut être appliqué [302]. Ces actions, telles l'*e-mailing*, le publipostage et le démarchage à l'aide de forces commerciales peuvent coûter beaucoup d'argent aux entreprises [301]. Pour pallier cet inconvénient, les décideurs ont besoin d'acquérir des connaissances pertinentes pour choisir les canaux de communication les plus efficaces et éviter continuellement une surabondance d'actions commerciales. Le CRM est devenu une stratégie pour les entreprises s'articulant autour d'un même concept [235] : « le client au centre de toutes les préoccupations » (cf. figure 2.5).

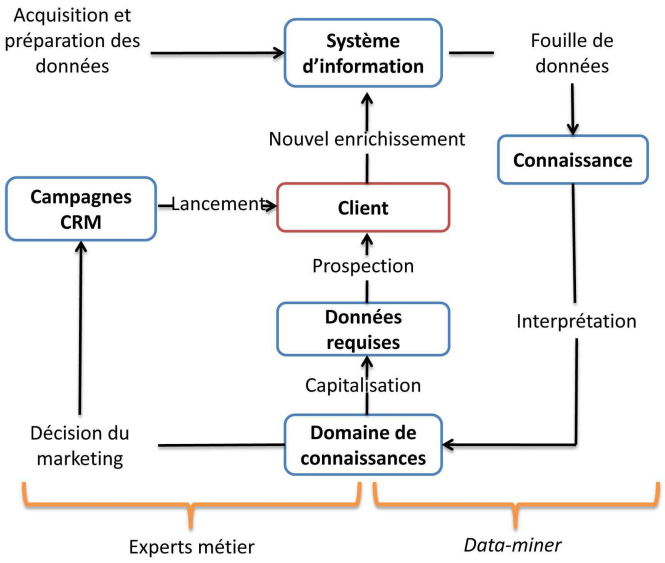


FIGURE 2.5 : Un système d'information évolutif centré sur le client [235]

La fouille de données permet de détecter des comportements clients implicites. Le CRM permet de déclencher des campagnes marketing en accord avec ces comportements. Ces deux derniers peuvent être combinés, aidant ainsi les experts métier à améliorer les relations et les interactions avec leurs clients [179, 242]. Une approche souvent utilisée consiste à trier les clients par probabilité estimée de réponse aux actions marketing et de comparer les classements en utilisant une courbe *lift* ou l'aire sous la courbe ROC [5] [136, 183]. De nombreux chercheurs ont abordé le problème du marketing direct comme un problème de classification [170, 171]. Néanmoins, les décideurs ont tendance à se concentrer sur la modélisation et l'interprétation des modèles, mais rarement sur l'action et l'estimation du retour sur investissement [255].

5. *Receiver Operating Characteristics.*

Anderson et al. [9] identifient dans le domaine de la gestion commerciale, l'organisation et la structuration requises pour un CRM efficace, les méthodes de fouille de données utilisées et les stratégies à mettre en place. L'étude [9] a été réalisée sur un grand nombre d'articles stockés sous forme de fichiers texte dans *QDA Miner*, un outil d'analyse qualitative pour répondre à trois besoins : i) l'organisation et l'infrastructure (structure de données, systèmes d'organisation, technologie, et accessibilité des données), ii) les objectifs des détaillants (stratégies axées sur le marketing, le service client, la compréhension des clients à travers l'analyse des données, la prospection et la rétention par le biais de programmes de fidélisation), iii) les outils d'exploration de données (efforts marketing et analyse de la clientèle). Les résultats donnent un aperçu des défis auxquels font face les entreprises pour axer leur stratégie sur le client.

Ngai et al. [205] présentent une classification thématisée des applications de fouille de données pour le CRM. Neuf cent articles ont été identifiés et examinés pour juger de la pertinence de l'application de techniques d'exploration de données pour le CRM. Les articles ont été classés suivant quatre dimensions : l'*identification*, l'*attraction*, la *fidélisation* et le *développement* de la valeur client, et sept méthodes de fouille de données [6, 99, 277]. Une classification des techniques de fouille de données pour le CRM est proposée en figure 2.6. Ce schéma a été élaboré à l'aide des recherches menées par Swift et al. [267], Sheth et al. [213], Ahmed et al. [6], Carrier et Povel [99], Mitra et al. [195] et Shaw et al. [255].

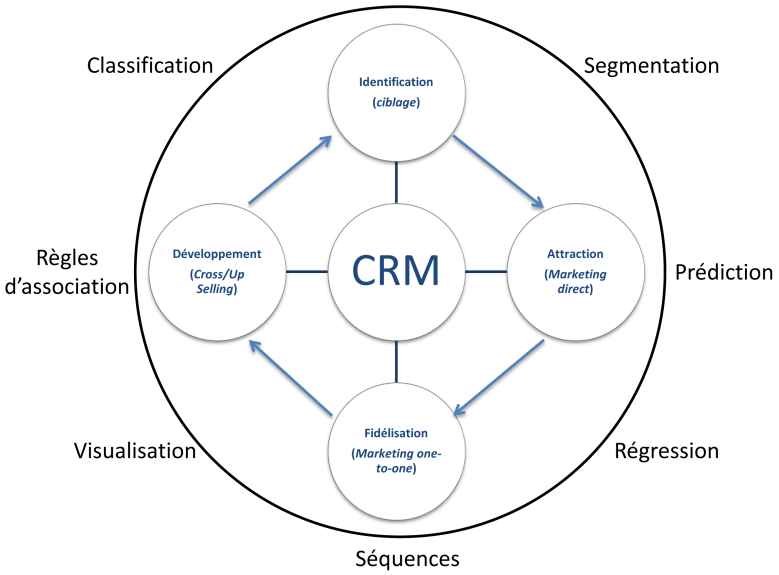


FIGURE 2.6 : Classification des techniques de fouille de données pour le CRM [205]

Les résultats soulignent que la fidélisation et le développement de la valeur client sont les domaines les plus représentatifs de l'application de la fouille de données pour le CRM. Les techniques de classification, d'extraction de règles d'association et de régression s'avèrent les plus couramment utilisées dans ces domaines.

2.1.6 De l'actionnabilité au retour sur investissement

Afin d'améliorer la relation client, les décideurs doivent connaître les mesures à actionner pour satisfaire les besoins économiques des entreprises (cf. tableau 2.1) [29]. Davenport et Harris [81] illustrent la manière dont les entreprises *leaders* de leur marché sont devenues de véritables « concurrents analytiques ».

Besoin économique	**Exemple d'action**	**Référence**
Prospection	Identification des prospects les plus susceptibles de devenir clients ou des anciens clients susceptibles de le redevenir.	Jiang et al. [142] Bitran et Mondschein [31]
Satisfaction	Promotions, prix et services. Personnalisation des pages de site Web en fonction du profil de chaque internaute.	Vanhoof et al. [284]
Interaction	Adaptation des canaux de communication aux segments de clients.	Bazzan [22]
Fidélisation	Optimisation de la tarification. Recommandations personnalisées.	Barwise et Farley [17]
Optimisation de la valeur client	Calcul de la rentabilité et de la *life time value* des clients. Identification des clients rentables et concentration des efforts commerciaux.	Venkatesan et Kumar [286]
Réduction des coûts	Réduction du nombre de clients ciblés. Meilleur taux de réponse lors des campagnes marketing.	Kohavi [152]
Gains de parts de marchés	Associations et recommandations produits (ventes croisées ou montées en gammes). Identification des clients susceptibles de partir à la concurrence. Détermination des meilleures implantations en fonction de leur localisation et des chiffres d'affaires générés.	Apte [10]

TABLEAU 2.1 : Les actions à déclencher face aux besoins économiques des entreprises

DÉFINITION 1

La notion de retour sur investissement [34] (RSI ou ROI pour Return on Investment), parfois appelé rendement, taux de rendement ou taux de profit, désigne un ratio financier qui mesure le montant d'argent gagné ou perdu par rapport à la somme initialement investie. Le ROI peut être défini de la manière suivante :

$$ROI = \frac{V_f - V_i}{V_i} = \frac{V_f}{V_i} - 1 \tag{2.1}$$

où V_i est la valeur initiale et V_f la valeur finale de l'investissement.

EXEMPLE 1 En fouille de données, le ROI peut fluctuer en fonction de l'optimisation de critères tels que l'augmentation des taux de réponses lors d'actions marketing, l'augmentation de la productivité des commerciaux, la meilleure répartition des ressources et la réduction des impayés.

Yan et al. [300] proposent un nouvel algorithme utilisant une descente de gradient pour optimiser directement les contraintes budgétaires des experts et ainsi maximiser le ROI. Hahn et al. [113] présentent une méthodologie pilotée par le ROI des sites Internet de commerce électronique. Breur [38] propose un test complet d'évaluation de la contribution des modèles de fouille de données pour le ROI. Enfin, Piton et al. [223] proposent une méthodologie d'application de méthodes de fouille de données pour optimiser le ROI des opérations commerciales.

Le ROI est souvent difficile à évaluer car les gains provenant de la mise en œuvre de la fouille de données ne sont pas toujours isolés de ceux qui proviennent d'une bonne communication, d'un marketing efficace et de commerciaux motivés [71]. Par exemple, une banque, après avoir déterminé pour chaque client son score de risque, son score d'appétence au crédit à la consommation et sa capacité de remboursement mensuel, a envoyé à chacun de ses clients ayant de bonnes notes une proposition personnalisée de crédit [276]. Les résultats ont été très supérieurs aux résultats habituels, de façon quantitative (augmentation du taux de souscription) et qualitative (appréciation des commerciaux). Dès lors, quelle est la part d'amélioration due à la qualité du ciblage et celle due à la personnalisation du mailing ? L'essentiel pour le ROI n'est pas de disposer des meilleurs outils de fouille de données mais de « savoir les insérer dans une démarche intégrée de marketing de bases de données » [275]. La fouille de données n'est qu'une brique du marketing de bases de données, parmi d'autres éléments tels que la communication marketing, le choix des argumentaires commerciaux, la forme des courriers, l'existence d'une relance téléphonique ou l'adaptation des processus : passer du marketing « produit » au marketing « client ».

Malgré ce constat, une approche simpliste existe pour mesurer un ROI : (i) élaborer un échantillon aléatoire de clients, (ii) traiter tous les clients de la même manière (canal de communication, support, relance, etc.) et (iii) comparer les résultats à l'issue de la campagne. Une comparaison possible du ciblage traditionnel et du ciblage à l'aide de techniques de fouille de données est présentée dans le tableau 2.2.

		Traditionnel	**Fouille de données**
A	Nombre de clients ciblés	30 000	15 000
B	Coût de chaque *mailing*	1 €	1 €
C	Coût de chaque relance téléphonique	5 €	5 €
D	Coût total $(A \times (B + C))$	180 000 €	90 00 €
E	Nombre de nouvelles souscriptions	1 000	1 500
F	Taux de souscription (E/A)	3,33 %	10 %
G	Coût par souscription	180 €	60 €
H	Chiffre d'affaires annuel par souscription	150 €	175 €
I	CA total annuel $(H \times E)$	150 000 €	262 500 €
	ROI $(= I/D)$	83 %	292 %

TABLEAU 2.2 : ROI : ciblage traditionnel vs ciblage par fouille de données [275]

Enfin, pour optimiser le calcul du ROI, il est conseillé de défalquer les coûts afférents aux logiciels et la masse salariale des *data miners*.

2.2 Extraction de connaissances actionnables

2.2.1 Techniques de fouille de données pour l'AKD

Dans cette partie, nous présentons l'utilisation actuelle des techniques de fouille de données pour l'extraction de connaissances actionnables (ou AKD pour *Actionable Knowledge Discovery*). Lobur et al. [178] proposent une synthèse des tendances actuelles de l'extraction de connaissances actionnables. Cao et al. [63] formalisent quatre types de découvertes de connaissances actionnables :

1. **PA-AKD** : *Post-Analysis-based Actionable Knowledge Discovery* : l'extraction et l'optimisation du modèle à l'aide des experts métier ;
2. **UI-AKD** : *Unified Interestingness-based Actionable Knowledge Discovery* : le développement de mesures d'intérêts, identifiant et décrivant les besoins techniques et métier ;
3. **CM-AKD** : *Combined Mining-based Actionable Knowledge Discovery* : l'extraction en plusieurs itérations des modèles dans l'objectif de choisir le meilleur ;
4. **MSCM-AKD** : *Multi-Source Combined Mining-based Actionable Knowledge Discovery* : l'utilisation de toutes les sources de connaissances techniques ou métier pour améliorer les modèles.

Nous illustrons dans les sous parties suivantes des exemples d'applications de techniques de fouille de données courantes pour l'extraction de connaissances actionnables : le *clustering*, les arbres de décision, le *scoring* et les réseaux bayésiens.

2.2.1.1 *Clustering*

Adams et al. [1] comparent, dans le contexte des crédits à la consommation, les différents domaines d'activité impactés par les méthodes de fouille de données face aux méthodes traditionnelles. L'auteur se penche sur l'analyse des comportements des clients d'une banque britannique en analysant l'évolution du statut des comptes courants au cours d'une période de douze mois. Les méthodes de *clustering* peuvent être utilisées pour définir de grandes catégories de comportement clients permettant ainsi de déclencher des actions pour une meilleure maîtrise du statut de leurs comptes respectifs.
Kelly [146] souligne l'apport de segmenter une base clients à l'aide de données statiques (style de vie, préférences, goûts, etc.) et de données dynamiques (achats, fidélité, rentabilité, etc.). Les résultats de la segmentation sont ensuite confrontés aux segments traditionnels créés par les experts métier de l'entreprise, trouvant ainsi le meilleur compromis avant le déclenchement d'une action commerciale.
Thonnard et Dacier [271] proposent une méthodologie multidimensionnelle de découverte de connaissances actionnables basée sur des techniques de *clustering* pour la détection de menaces sur Internet, prenant en compte les conseils des experts et les contraintes du domaine. L'objectif est double : i) développer des indicateurs pour

évaluer la prévalence de certaines activités malveillantes et ii) obtenir un aperçu des nouvelles attaques émergentes afin d'améliorer la compréhension des menaces par les experts métier.

2.2.1.2 Arbres de décisions

Yang et al. [302] présentent une technique d'extraction de connaissances actionnables à partir des arbres de décisions dans le domaine des télécommunications. Un algorithme de post-traitement des arbres de décisions, *Greedy-BSP*, est utilisé afin d'obtenir des actions associées au changement de valeurs de certaines variables. La première idée est d'extraire de la connaissance actionnable quand il n'y a pas de restriction du nombre d'actions (*unlimited-resource case*). La deuxième idée est de restreindre le nombre d'actions en donnant un coût à chaque déclenchement d'action. L'algorithme de base, *Leaf-Node Search*, permet de changer un client de feuille dans l'arbre de décision en fonction de probabilités maximisant le profit généré par client. En effet, changer un client de classe sous-entend que certains de ses attributs vont être modifiés. Pour changer les attributs d'un client, il faut mettre en place des actions qui ont un coût. Cependant, le client peut par la suite générer davantage de profit pour l'entreprise. De plus, certains attributs ne peuvent pas changer de valeurs (le *sexe* par exemple). Ces derniers sont appelés (*hard attributes*) et présentent des valeurs significatives dans la matrice des coûts. En revanche les attributs pouvant changer plus facilement de valeur (le *département d'origine* par exemple) sont appelés *soft attributes*. La matrice des coûts est rarement symétrique. En effet, le sens d'un changement de classe peut opposer des coûts complètement différents. Par exemple, le coût de transformation d'un client prospect à un client fidèle peut être élevé. En revanche, le coût de transformation d'un client fidèle à un client « endormi » peut être très faible, sinon gratuit.
Gunnarsson et al. [110] illustrent une étude de cas dans laquelle l'arbre de décision est utilisé comme un outil de prévention de désabonnement pour un grand quotidien du *Midwest USA*. Les connaissances extraites tout au long du processus de fouille offrent un partage d'information pour l'entreprise et une optimisation de sa prise de décision (*data-driven decision making*).

2.2.1.3 *Scoring*

Pour illustrer concrètement l'apport des connaissances actionnables pour les entreprises, nous décrivons dans cette partie l'application d'une importante branche de la fouille de données : le calcul du score ou *scoring*. Par son ancienneté et son universalité, on peut voir le *scoring* comme l'archétype des applications de fouille de données en entreprise. Les techniques prédictives les plus utilisées pour le *scoring* sont la régression logistique, l'analyse discriminante et les arbres de décision, parfois les SVM (*Support Vector Machine*) ou le classifieur bayésien naïf. Généralement, le *scoring* se base sur la régression linéaire permettant de mettre en relation une variable continue Y expliquée avec une variable explicative X continue. On suppose

communément que les valeurs $x_1, ..., x_n$ de X sont contrôlées et sans erreur de mesure, et on observe les valeurs correspondantes $y_1, ..., y_n$ de Y. On suppose que X et Y ne sont pas indépendantes et que la connaissance de X permet d'améliorer la connaissance de Y. Évidemment, savoir que $X = x$ permet rarement de connaître la valeur exacte de Y, mais la valeur moyenne $E(Y|X = x)$, c'est-à-dire l'espérance conditionnelle de Y sachant que $X = x$. Plus précisément, le postulat de base est que $E(Y|X = x)$ est une fonction linéaire de x, se traduisant par :

$$E(y_i) = \alpha + \beta x_i \ \forall_{i=1,...,n} \tag{2.2}$$

Types de scores
De nombreux types de scores sont utilisés au sein des entreprises pour mener des actions généralement profitables [235, 275] :

- Le score d'*appétence* (de *propension* ou d'*affinité*) mesure la probabilité pour un client d'être intéressé par un produit ou un service ;
- Le score de *risque* mesure la probabilité pour un client de rencontrer un incident (de paiement ou de remboursement par exemple) ;
- Le croisement des deux premiers scores est parfois appelé score de *pré-acceptation*, car il fournit une cible de clients à qui proposer des offres cibles comme par exemple des offres pré-acceptées de crédit ;
- Le score d'*octroi* ou d'*acceptation* est un score de risque calculé pour un nouveau client ou un client de faible activité avec l'entreprise ;
- Le score de *recouvrement* évalue le montant susceptible d'être récupéré sur un compte ou un crédit au contentieux, et peut suggérer des actions ;
- Le score d'*attrition* (*churning*) mesure la probabilité pour un client de quitter l'entreprise pour la concurrence [294].

Scores actionnables
Finlay [93] apporte des éléments de comparaison liant le potentiel client à ses risques de remboursement, permettant ainsi de trouver le meilleur compromis entre développement de la valeur client et risque encourus pour l'entreprise. Zhu et al. [313] utilisent des scores de crédit permettant de mesurer la solvabilité d'individus. Pour y parvenir, ils définissent les concepts de suffisance (*sufficiency*) et d'extranéité (*extraneousness*) pour étudier les conditions dans lesquelles les résultats de notation peuvent être améliorés à l'aide de combinaisons de scores individuels. Le concept de suffisance est utilisé pour identifier des scores qui sont dominants. L'extranéité quant à elle est utilisée pour déterminer si un score fournit des informations utiles et différentes des autres scores. Une mesure d'utilité axée sur le profit permet d'évaluer la performance des différents scores. Liu et Schumann [177] présentent une étude empirique de quatre méthodes d'apprentissage. Ces méthodes fournissent une technique automatisée d'exploration de données pour réduire l'espace des dimensions.

L'étude illustre comment quatre algorithmes : *ReliefF*, *Correlation-based*, *Consistency-based* et *Wrapper* aident à améliorer trois aspects de la performance des modèles d'évaluation : la simplicité, la vitesse de convergence et la précision. Les expérimentations sont menées sur des données réelles au moyen de quatre algorithmes de classification : *Model Tree (M5)*, *Neural Network*, *Logistic Regression* et *K-Nearest-Neighbours*. Enfin, Cumby et al. [77] décrivent un prototype de prédiction de listes d'achats dans le domaine de la vente au détail. Ils réalisent un score distinct pour chaque client à partir de données d'historiques d'achats. Les résultats soulignent une prédiction des listes d'achats précise et robuste. Pour les experts métier, le gain a été autant quantitatif (augmentation du chiffre d'affaires de près de 11 %) que qualitatif (fidélisation et satisfaction client).

2.2.1.4 Réseaux Bayésiens

Les exemples présentés dans cette sous section sont inspirés de Naïm et al. [202].

- L'une des applications qui fait référence pour l'utilisation des réseaux bayésiens pour la fouille de données est le système de détection de fraude mis en production par la société de télécommunications ATT [87]. L'application développée vise deux objectifs : i) premièrement, détecter au niveau des clients ou des appels, un risque élevé de non recouvrement et, deuxièmement, décider les actions à effectuer en fonction de ce niveau de risque.

- L'application *Vista* a été développée par la NASA en collaboration avec la société *Knowledge Industries* [130]. Cette application est fondée sur la recherche d'un compromis entre le temps nécessaire pour prendre une décision, qui augmente avec le nombre d'informations à analyser, et le temps disponible pour prendre cette décision, qui peut être court si le système concerné évolue rapidement. Cet arbitrage est particulièrement sensible dans *Vista* qui est le suivi des moteurs de positionnement orbital de la navette spatiale américaine.

- La société Ricoh a été l'une des pionnières de l'utilisation des réseaux bayésiens en développant un système d'assistance aux opérateurs chargés d'intervenir sur des copieurs en panne [119]. L'approche utilisée pour construire ce système appelé *Fixit* est relativement originale. En fonction des symptômes décrits par l'utilisateur, *Fixit* recherche les causes de pannes possibles, et présente directement à l'utilisateur une aide documentaire adaptée.

- Le projet *Lumière* [131], centré sur la construction et l'intégration de modèles bayésiens pour l'aide à l'utilisateur, a conduit à définir le produit *Office Assistant*. Plus récemment, les réseaux bayésiens ont trouvé une nouvelle application dans le domaine informatique : l'*antispam*, c'est-à-dire le filtrage des emails non sollicités. Le groupe DTAS de *Microsoft* détecte les emails les plus pertinents. Une solution appelée *Mobile Manager* identifie les messages les plus importants et en informe le destinataire par une notification sur son mobile.

Dans la section suivante, nous verrons que des méthodologies prennent en compte les contraintes du domaine pour l'extraction de connaissances actionnables.

2.2.2 *Domain Driven Actionable Knowledge Delivery*

2.2.2.1 Concept

La connaissance extraite des différents modèles doit permettre aux experts métier de comprendre le comportement de leurs clients et de prendre des décisions à bon escient (paradigme de l'*actionable knowledge* [52, 104]). La réussite de tout processus d'extraction de connaissances nécessite une étroite collaboration avec les experts du domaine, à toutes les étapes du processus [260]. Le D^{3}M [61, 62] pour *Domain Driven Actionable Knowledge Delivery* ou *Domain Driven Data Mining* [54, 59] prend en compte cette nécessité en offrant des méthodologies, des techniques, et des outils de fouille qui s'adaptent aux contraintes des entreprises et qui visent à promouvoir le changement de paradigme de l'extraction de motifs cachés à partir des données à la découverte de connaissances actionnables [53]. À cette fin, le D^{3}M doit impliquer et intégrer l'intelligence des données (*data intelligence*), l'intelligence du domaine (*domain intelligence*), l'intelligence organisationnelle (*network intelligence*), l'intelligence humaine (*human Intelligence*), l'intelligence sociale (*social Intelligence*) et une intelligence métasynthétisée (*intelligence metasynthesis*). Les décideurs métier détiennent le droit de qualifier de « bonnes » ou « mauvaises » les connaissances extraites. Ainsi, le D^{3}M a pour objectif de découvrir de la connaissance actionnable en fonction des contraintes réelles des entreprises. Ce type de démarche a toute son importance dans la recherche et le développement des futures méthodologies et infrastructures de fouille de données [60].
Le tableau 2.3 permet de comparer l'approche D^{3}M orientée domaine (*Domain Driven*) à une approche traditionnelle orientée données (*Data-Centered*).

	Data-Centered	***Domain Driven***
Finalité	Élaboration d'approches novatrices	Résolution de problèmes métier
But	Démontrer et justifier l'utilisation de nouveaux algorithmes pour découvrir des connaissances	Découvrir des connaissances cachées répondant à des besoins réels métier
Cible	Algorithmes	Problèmes métier
Élément de fouille	Données	Données et domaine
Jeu de données	L'ensemble des données	Données sous contraintes
Modèle	Prédéfini	Personnalisé
Processus	Automatique	Intervention des experts métier
Performance	Précision et optimisation	Actionnabilité
Évaluation	Mesures techniques	Interprétation des décideurs
Livrables	Modèles cachés	Actions métier

TABLEAU 2.3 : *Data-Centered* vs *Domain Driven Data Mining* [60]

La figure 2.7 illustre la carte de concepts de la méthodologie D^{3}M. Cao a décomposé la carte de concepts en quatre couches séquentielles et complémentaires, de la plus extérieure à la plus centrale :

1. Couche *Specific domain* : les domaines ;
2. Couche *Fundamental issues* : actionnabilité, opérabilité, utilisabilité, etc. ;
3. Couche *D*^{3}*M theoritical foundations* : mathématiques, sciences, etc. ;
4. Couche *D*^{3}*M techniques and tools* : intelligences et représentations.

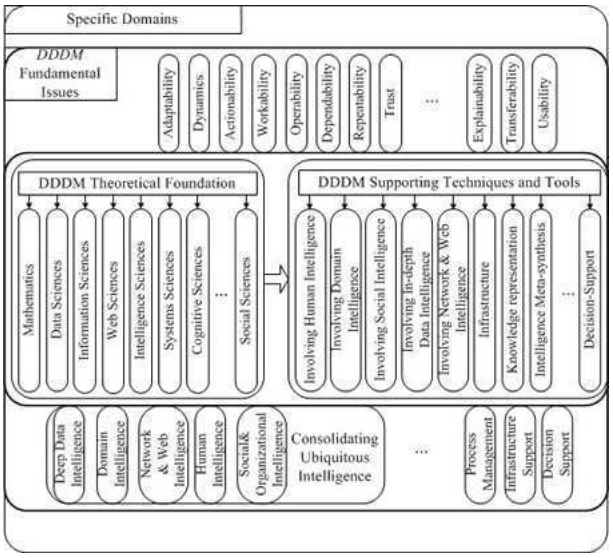


FIGURE 2.7 : Carte de concept de la méthodologie D^{3}M [60]

2.2.2.2 Méthodologies

Plusieurs méthodologies ont été proposées pour satisfaire l'approche D^{3}M.

CRISP-DM [256], qui signifie *CRoss-Industry Standard Process for Data Mining*, est une méthodologie mise au point en 1996 et approuvée dans différents contextes à l'aide d'experts métier. Cette méthodologie décrit les phases typiques d'extraction de connaissances à partir des données et les tâches à accomplir pour y parvenir. Cette méthodologie offre un aperçu du cycle de vie du processus de fouille de données. Les principales étapes sont les suivantes :

1. **Compréhension de la problématique métier** : description du contexte, détermination des objectifs et des critères de réussite et inventaire des ressources disponibles ;
2. **Compréhension des données** : collecte, description, exploration et contrôle de la qualité des données ;
3. **Préparation des données** : sélection des données nécessaires à l'étude, nettoyage des données, construction et consolidation ;

4. **Modélisation** : choix de la ou des techniques pouvant répondre à la problématique et sélection du meilleur modèle ;

5. **Évaluation et test** : application du modèle sur l'échantillon de test et vérification du classement ;

6. **Déploiement** : création des règles de déploiement, analyse des résultats, rapport d'étude et préconisation d'actions.

DDID-PD, qui signifie *Domain-Driven In-Depth Pattern Discovery* [59, 60] est une méthodologie pour l'extraction de connaissances reposant sur les besoins réels des experts métier (cf. figure 2.8). Il s'agit de prendre en compte continuellement l'extraction sous contrainte du domaine, la coopération avec les experts du métier et l'itération récurrente de la méthodologie.

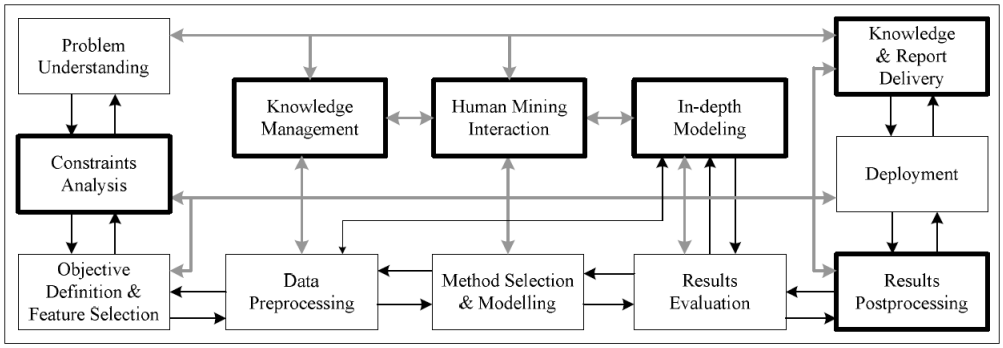


FIGURE 2.8 : Processus de l'approche DDID-PD [59]

Quatre étapes guidées par les utilisateurs sont nécessaires pour une découverte de connaissances applicables aux problématiques industrielles :

1. *Including constraint mining* : l'objectif de cette étape est d'inclure des contraintes fonctionnelles, non fonctionnelles et environnementales dans l'algorithme ;

2. *In-depth mining* : généralement de nombreux algorithmes ne prennent pas en compte le contexte *business* de l'entreprise, nommés les *generic patterns*. C'est pourquoi, il est intéressant de mesurer l'actionnabilité d'un modèle à travers différentes mesures métier (cf. section 2.3.4) ;

3. *Human Cooperated Mining* : cette étape représente la coopération entre l'intelligence humaine (qualitative) et l'intelligence extraite (quantitative) ;

4. *Loop-closed mining* : le processus d'extraction de connaissance actionnable est itératif (évaluations, raffinements, nouvelles hypothèses, etc.) avant d'atteindre la connaissance actionnable pour les experts métier.

Une fois le modèle réalisé, la méthodologie *DDID-PD* offrent différents points de vues pour mesurer l'actionnabilité du modèle :

- *Technical interestingness* : mise en place d'indicateurs objectifs (souvent extraits des algorithmes) pour juger de la qualité de la connaissance. Par exemple, dans le cadre des règles d'association, une règle sera extraite si elle satisfait un minimum de support et un minimum de confiance (ces deux indicateurs étant préalablement ajustés en fonction de la connaissance des experts métier) ;
- *Business interestingness* : mise en place d'indicateurs métier pour satisfaire au plus près les contraintes du domaine. Par exemple, un seuil minimum de ROI peut être fixé pour l'acceptation d'un modèle par les experts ;
- *Actionability of a pattern* : cette notion indique à quel degré les deux indicateurs précédents sont satisfaisants. Dans le cas où les deux indicateurs sont satisfaisants, le modèle est *actionnable*. En revanche, le modèle peut être intéressant pour les experts métier et non pour les *data miners*.

2.2.2.3 Applications

Ces dernières années ont été fortement impactées par l'ECD et plus particulièrement par l'AKD. En effet, les applications de fouille de données ont subi une transformation importante [178], influencée par des facteurs extérieurs tels que la croissance d'Internet. D'autres domaines d'application ont également été impactés. Par exemple, le monde de la finance [55, 179] ou de la sécurité [57]. De nouveaux algorithmes ont été proposés, suggérant ainsi des actions pour aider les experts à transformer le comportement de leurs clients pour accroître la fidélité tout en maximisant le profit de l'entreprise [302], ou encore à travers des approches hiérarchiques d'actions [278]. La fouille de données impacte de plus en plus en entreprise la prise de décision des experts métier [114].

2.3 Actionnabilité des règles d'association

2.3.1 Terminologie et notations

Considérons un ensemble T de n transactions décrites par un ensemble I de variables qualitatives ou quantitatives. T est stocké sous forme de table dans une base de données relationnelle. L'ensemble $I = \{i_1, i_2, \ldots i_m\}$ représente un ensemble de m attributs décrivant les n transactions de la base de données. Ces variables sont appelées des *items* dans la terminologie des règles d'association. L'ensemble $T = \{t_1, t_2, \ldots t_n\}$ représente un ensemble de transactions, une transaction étant un sous ensemble de I. Une conjonction d'items $Z = \{i_1, i_2, \ldots i_k\}$ non vide de I est appelé un *itemset*. Le nombre d'items k de l'ensemble Z constitue sa longueur. Un *itemset* de longueur k est nommé *k-itemset*.
Un ensemble d'items est dit **fréquent** si et seulement si il correspond à un motif fréquent dans la base de transactions. Nous définissons le support d'un motif comme étant le nombre de transactions dans T contenant ce motif divisé par le cardinal de T, nommé $card(T)$. Un seuil minimal de support $minSup$ est fixé à partir duquel un ensemble d'items est dit fréquent.

2.3.2 Règles d'association

Une règle d'association [32] est une liaison orientée entre variables booléennes de la forme $X \rightarrow Y$, où X et Y sont des itemsets ou négations d'itemsets qui n'ont pas d'items en commun. Elle traduit la tendance de Y à être vrai quand X est vrai, et peut s'énoncer de la manière suivante : « si un individu vérifie X alors il vérifie sûrement Y ». X est appelé prémisse et Y conclusion de la règle. Les exemples d'une règle sont les individus qui vérifient la prémisse et la conclusion. En revanche, les contre-exemples sont les individus qui vérifient la prémisse et non la conclusion. À partir de deux variables X et Y, il est possible de construire huit règles différentes (cf. tableau 2.4).

$1 : X \rightarrow Y$	$2 : X \rightarrow \overline{Y}$	$3 : \overline{X} \rightarrow Y$	$4 : \overline{X} \rightarrow \overline{Y}$
$5 : Y \rightarrow X$	$6 : Y \rightarrow \overline{X}$	$7 : \overline{Y} \rightarrow X$	$8 : \overline{Y} \rightarrow \overline{X}$

TABLEAU 2.4 : Règles d'association générées à partir de deux variables X et Y

De plus, pour une règle $X \rightarrow Y$, nous utilisons les notations suivantes :

- $X \rightarrow \overline{Y}$ est la règle contraire, $Y \rightarrow X$ la réciproque et $\overline{Y} \rightarrow \overline{X}$ la contraposée ;
- $n = |T|$, le nombre de transactions ou enregistrements ;
- $n_x = |X|$, le nombre de transactions satisfaisant X ;
- $n_y = |Y|$, le nombre de transactions satisfaisant Y ;
- $n_{xy} = |X \cap Y|$, le nombre de transactions satisfaisant à la fois X et Y ;
- $n_{x\overline{y}} : |X \cap \overline{Y}|$, le nombre de transactions satisfaisant X mais pas Y.

Soit x une variable booléenne qui est un itemset ou une négation d'itemset. La variable $\overline{x}$ est la négation de x. Nous notons X l'ensemble des transactions de T vérifiant x et n_x le cardinal de X. Le complémentaire de X dans T est $\overline{X}$ de cardinal $n_{\overline{x}}$. La probabilité de l'événement x *est vrai* est notée $P(x)$. Elle est estimée par la fréquence empirique (estimateur du maximum de vraisemblance) : $P(x) = \frac{n_x}{n}$.

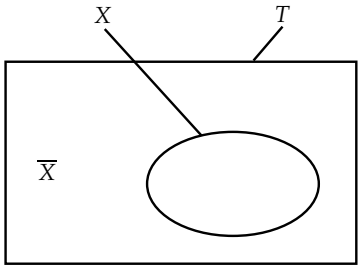


FIGURE 2.9 : Diagramme de Venn de l'ensemble X dans celui des transactions T [32]

2.3.3 Algorithmes

La technique d'extraction de règles d'association introduite par Agrawal en 1993 [4] est considérée comme l'une des techniques les plus importantes dans la découverte de connaissances dans les bases de données [90]. *AIS* [4] fut l'un des premiers algorithmes écrits pour les centrales d'achats. Par la suite, en 1994, l'algorithme *Apriori* [5] est proposé. Deux étapes principales composent ces algorithmes :

1. La recherche des itemsets fréquents ;
2. La génération des règles d'association.

Différentes études soulignent que ce découpage correspond à la manière optimale face à l'explosion combinatoire de l'algorithme. Ainsi, la recherche des itemsets fréquents sur un ensemble m d'attributs binaires est un espace de dimension 2^m. L'étape de génération des règles à partir d'un itemset fréquent de largeur k peut engendrer la génération de $2^k - 2$ règles.
Depuis l'algorithme de référence d'Agrawal et Srikant [5], de nombreux algorithmes ont été développés et sont synthétisés dans Hipp et al. [129]. Différents algorithmes ont également été développés pour réduire le nombre d'itemsets en générant des itemsets *fermés*, *maximaux* ou *optimaux* [46, 163, 304]. D'autres algorithmes ont été développés pour réduire le nombre de règles [214, 305, 306]. Enfin, différentes méthodes complémentaires de post-traitement ont été proposées telles que l'élagage (*pruning*), les résumés (*summarizing*) ou les regroupements (*grouping*) [13, 15, 174, 207, 273].

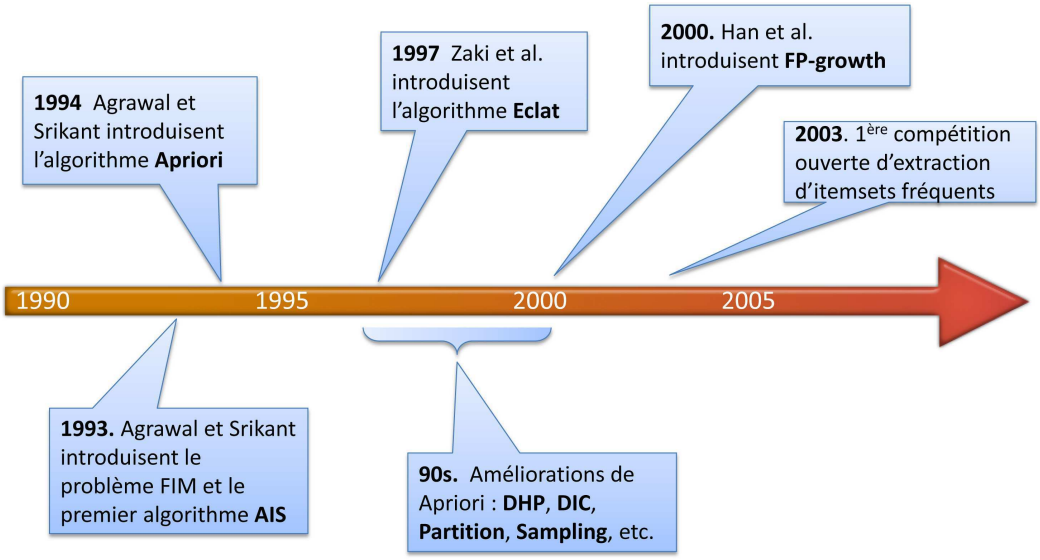


FIGURE 2.10 : Historique des algorithmes d'extraction d'itemsets fréquents

Extraction des itemsets fréquents (ou FIM pour *Frequent Itemset Mining*)
Les algorithmes d'extraction des itemsets fréquents considèrent un ensemble d'itemsets d'une taille donnée lors de chaque itération, c'est-à-dire un ensemble d'itemsets d'un « niveau » du treillis des itemsets. Ces algorithmes limitent le nombre d'itemsets candidats considérés, en les générant à partir des itemsets fréquents de l'itération précédente : « tous les sur-ensembles d'un itemset infréquent sont infréquents et tous les sous-ensembles d'un itemset fréquent sont fréquents » [5, 181].
Les algorithmes **Apriori** [5] et **OCD** [181] réalisent un nombre de balayages du contexte égal à la taille des plus longs itemsets fréquents. L'algorithme **Partition** [247] autorise la parallélisation du processus d'extraction. L'algorithme **DIC** [41] réduit le nombre de balayage du contexte en considérant les itemsets de plusieurs tailles différentes lors de chaque itération. Les algorithmes Partition et DIC entraînent un coût supplémentaire en temps CPU [6] par rapport aux algorithmes Apriori et OCD, dû à l'augmentation du nombre d'itemsets candidats testés.

Extraction des itemsets fréquents maximaux
Ces algorithmes se basent sur le fait que les itemsets fréquents maximaux, c'est-à-dire les itemsets dont tous les sur-ensembles sont infréquents, forment une bordure au dessous de laquelle tous les itemsets sont fréquents. L'extraction des itemsets fréquents maximaux est réalisée par une exploration itérative du treillis des itemsets fréquents. À partir des itemsets fréquents maximaux, tous les itemsets fréquents sont dérivés et leurs supports sont déterminés en réalisant un balayage du contexte. Quatre algorithmes basés sur cette approche ont été proposés : **Pincer-Search** [166], **MaxClique**, **MaxEclat** [307], et **Max-Miner** [20]. Ces algorithmes permettent de réduire le nombre d'itérations et donc de diminuer le nombre de balayages du contexte et d'opérations CPU réalisées.

Extraction des itemsets fréquents fermés
Les itemsets fréquents fermés [214] sont définis avec la fermeture de la connexion de Galois d'une relation binaire finie. Tous les itemsets fréquents et leurs supports, et donc toutes les règles d'association ainsi que leurs support et confiance, peuvent être déduits efficacement, sans accéder au jeu de données à partir des itemsets fréquents fermés. Les algorithmes **Close** et **A-Close** [214] sont des algorithmes d'extraction des itemsets fermés : ils considèrent un ensemble de générateurs candidats d'une taille donnée, et déterminent leurs support et fermeture en réalisant un balayage du contexte lors de chaque itération. L'algorithme **Close+** permet d'identifier les itemsets fréquents fermés et leurs générateurs parmi les itemsets fréquents, sans accéder au jeu de données. Notons que tous les itemsets et les itemsets fermés sont fréquents et que tous les itemsets fréquents maximaux sont fermés (cf. figure 2.11).
Han et al. [117] ont récemment classés les algorithmes d'extractions de règles d'association en trois catégories :

1. Les méthodes au format horizontal basées sur *Apriori* : par exemple *Apriori* [5] ;

6. *Central Processing Unit*

2. Les méthodes au format vertical basées sur *Apriori* : par exemple *CHARM* [306] ;
3. Les méthodes de projections pour le parcours de structures compressées : par exemples *FP-growth* et *FP-tree* [105].

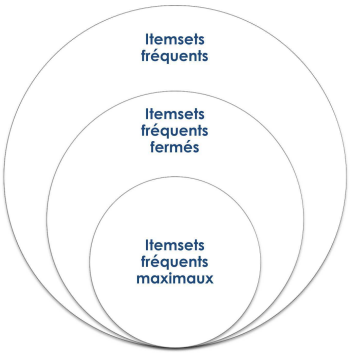


FIGURE 2.11 : Itemsets fréquents, fermés et maximaux

Le paramètre commun de ces trois approches est la détermination du seuil de support minimal assurant la génération des itemsets fréquents. Cette approche présente tout de même un inconvénient : un seuil de support trop faible peut conduire à la génération d'un nombre trop importants d'itemsets tandis qu'un seuil trop important peut minimiser le nombre de règles. Par conséquent, la recherche des *top-k itemsets fréquents fermés* a été proposée, de longueur minimale *min_l*, où *k* est un nombre déterminé par l'expert permettant de fouiller *k* itemsets fréquents fermés. *Top-k* fait référence aux *k* motifs fermés les plus fréquents et *min_l* la longueur minimale des motifs fermés. Pour répondre à cette approche, plusieurs algorithmes ont été proposés [280, 298], dont TFP (*Top-k Frequent Closed Itemsets*) [288] dans lequel toutes les transactions plus courtes que *min_l* sont exclues de la recherche et le support minimal est déterminé dynamiquement lors de la construction du *FP-tree*, permettant ainsi d'élaguer le treillis avant l'extraction.

2.3.4 Mesures de qualité

Les mesures de qualité ou d'intérêts des règles d'association aident les experts à trouver des connaissances intéressantes dans de grands volumes de règles [39] (cf. figure 2.13). Ces mesures permettent d'évaluer les règles [21], de sélectionner celles respectant un seuil de qualité minimale [268] et de les ordonner des plus actionnables au plus irréalisables [33] (cf. figure 2.13). Face aux nombreuses mesures d'intérêt présentes dans la littérature, Piatetsky-Shapiro et Matheus [221] proposent de les catégoriser de la manière suivante :

- Les mesures dites **objectives** (ou *data-oriented*) dépendantes des données ;
- Les mesures dites **subjectives** (ou *user-oriented*) prenant en compte les objectifs et les croyances des utilisateurs.

Cette classification a été revue par Silberchatz et Tuzilin [259] un peu plus tard introduisant une classification des mesures subjectives en mesures actionnables (*actionability*) et inattendues (*unexpectedness*). Différents critères discriminants [109] ont été utilisés pour la définition des mesures d'intérêts :

1. La **Concision** ou la longueur de la règle d'association ;
2. La **Généralité** ou le support de la règle d'association ;
3. La **Fiabilité** ou la confiance de la règle d'association ;
4. La **Validité** ou l'association généralité/fiabilité de la règle d'association ;
5. La **Particularité** ou la manière de gérer les données aberrantes ;
6. La **Diversité** ou la comparaison avec une distribution uniforme ;
7. La **Nouveauté** ou l'apport de connaissances non déductibles par les experts ;
8. L'**Inattendu** ou la contradiction avec les connaissances expertes ;
9. L'**Utilité** ou l'accord avec une fonction d'utilité ;
10. L'**Applicabilité** ou le déclenchement d'une action dans le domaine étudié.

Geng et Hamilton [98] proposent une catégorisation des mesures d'intérêts sur laquelle nous nous reposerons dans la suite de nos travaux (cf. figure 2.12).

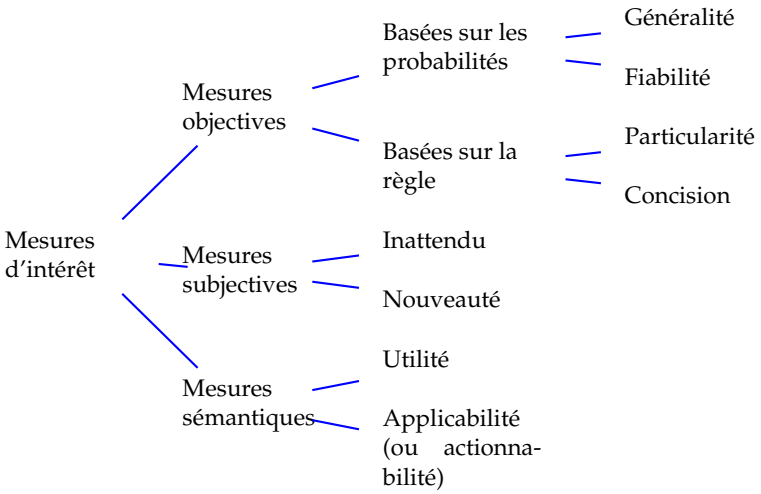


FIGURE 2.12 : Catégorisation des mesures d'intérêts des règles d'association [98]

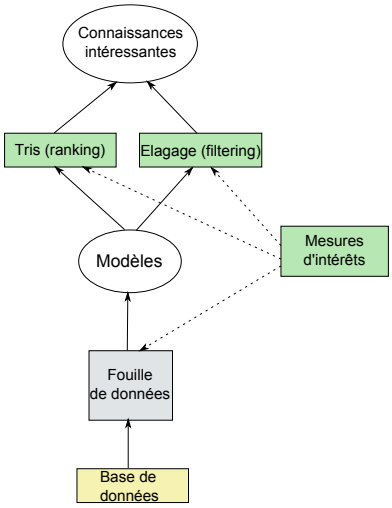


FIGURE 2.13 : Rôles des mesures d'intérêts dans le processus d'ECD [98]

2.3.4.1 Mesures objectives

Une méthode populaire pour évaluer l'intérêt des règles d'association consiste à quantifier cet intérêt à l'aide de mesures objectives. Définies à partir de la contingence des règles (caractéristiques des données), les mesures objectives permettent de classer les règles. Cependant, l'utilisation de mesures différentes peut engendrer un classement des règles fortement dissemblable [282]. Ces indices sont adaptés à des particularités de distribution de données ou à des attentes de l'utilisateur et le choix influe directement sur le jeu de règles découvert. La comparaison des indices se base en général sur deux types de critères :

- **Théoriques**, établissant un jeu de caractéristiques attendus et les confrontant avec les propriétés mathématiques des indices [33] ;
- **Expérimentaux**, appliquant ces indices sur des jeux de tests et éventuellement utilisant des outils de comparaison de règles d'association [268].

Les nombreuses comparaisons proposées par Piatetsky-Shapiro [219] et Tan et al. [268] montrent qu'aucune mesure ne respecte l'ensemble des critères proposés. Le choix de la mesure dépend fortement des résultats attendus et des paramètres du domaine. Par exemple, la distribution initiale des données entraîne souvent l'utilisation de mesures différentes [268]. Blanchard et al. [33] proposent une classification sémantique basée sur les déviations par rapport à l'indépendance et à l'équilibre et introduisent la qualité des règles d'association au travers de trois concepts : la **généralité** qui peut être calculée par le support et le support causal, la **puissance implicative** qui peut être mesurée par la confiance, l'indice de Loevinger ou la J-Mesure et la **significativité statistique** qui peut être mesurée par l'intensité implication. Les nombreuses mesures existantes et les propriétés de ces mesures ont ainsi suscité un grand nombre de travaux. Je renvoie le lecteur aux synthèses proposées par Lenca et al. [158, 159], Gras et al. [106] et Geng et Hamilton [98].

Support et confiance : les limites

DÉFINITION 2
Le support évalue la généralité de la règle et permet de réduire le nombre de règles à l'aide d'un élagage en fonction d'un seuil minSup. Cependant, l'extraction peut omettre la découverte de « pépites » car généralement, ces dernières, présentent un support faible pour une confiance élevée.

$$support(X \rightarrow Y) = \frac{n_{xy}}{n} \tag{2.3}$$

DÉFINITION 3
La confiance évalue la validité d'une règle et permet de réduire le nombre de règles à l'aide d'un élagage en fonction d'un seuil minConf. Cependant, cette mesure ne permet pas de détecter l'indépendance des variables. Elle estime la probabilité conditionnelle que la variable Y soit réalisée sachant que la variable X l'est.

$$confiance(X \rightarrow Y) = \frac{n_{xy}}{n_x} \tag{2.4}$$

Le support et la confiance sont deux indices élémentaires, mais ils constituent les mesures les plus communément utilisées pour évaluer les règles d'association. Malheureusement, ces indices présentent certaines limites [21, 268] : l'énorme quantité de règles d'association limite l'utilité de la technique et la trivialité d'une grande partie des règles générées implique une réelle nécessité de proposer de nouvelles mesures.

Le *lift* (ou l'intérêt)

DÉFINITION 4
La mesure du lift a été introduite par Brin et al. en 1997 [40]. Le lift correspond à une mesure d'écart par rapport à l'indépendance. La mesure du lift met en évidence l'intérêt d'une règle d'association comme suit :

$$lift(X \rightarrow Y) = \frac{nn_{xy}}{n_x n_y} = \frac{confiance(X,Y)}{n_y} \tag{2.5}$$

EXEMPLE 2 Un *lift* égal à 2 signifie que le nombre d'exemples de la règle $X \rightarrow Y$ est 2 fois plus grand que celui attendu en cas d'indépendance, ce qui veut dire que le consommateur qui achète X a 2 fois plus de chance d'acheter Y qu'un consommateur lambda. Le *lift* étant symétrique, la situation inverse est vraie puisque les exemples de $X \rightarrow Y$ sont aussi ceux de $Y \rightarrow X$.

La sélection de règles de manière purement statistique (mesures objectives) se heurte à deux écueils principaux : la production d'un trop grand nombre de règles et l'élimination de règles intéressantes si des critères trop restrictifs sont appliqués. Pour y remédier, nous présentons les mesures d'intérêts subjectives dans la partie 2.3.4.2.

2.3.4.2 Mesures subjectives

Les mesures d'intérêts subjectives [137] prennent en compte les objectifs et les connaissances des experts. Il s'agit de méthodes supervisées de traitement des règles suite à leur extraction par des algorithmes classiques, prenant en compte les savoirs et les questionnements des décideurs.
Plusieurs approches ont été proposées dans la littérature. Liu et al. [173] définissent les mesures subjectives comme la distance séparant les règles des spécifications des experts. Sahar [243] quant à elle propose davantage une approche en interaction avec les experts en laissant la possibilité de marquer les règles intéressantes. Padmanabhan et Tuzhilin [209] présentent une recherche de règles venant contredire les croyances des experts. Par la suite, les travaux de Silberschatz et Tuzhilin [259] permettent de définir une mesure subjective en comparant les probabilités a priori et a posteriori. Finalement, une approche plus récente de Marinica et Guillet [182] présente une méthode basée sur les schémas de règles et les ontologies.
Liu et al [173] proposent deux critères subjectifs de sélection de règles : leur caractère inattendu (*unexpectedness*) et leur actionnabilité (*actionability*). L'inattendu caractérise le fait que les règles sont intéressantes si elles surprennent les experts métier, c'est-à-dire si elles représentent pour lui une nouveauté. L'actionnabilité mesure l'aptitude d'une règle à être exploitée par l'utilisateur dans le cadre d'une action précise et utile. Nous présentons dans le tableau 2.5 quelques mesures subjectives rencontrées dans la littérature ainsi que les références bibliographiques associées.

Mesure subjective	**Référence**
Generality	Srikant et Agrawal [264]
Surprisingness	Liu et al. [173]
Actionability	Shapiro et Matheus [221]
Unexpectedness	Silberschatz et Tuzhilin [260]
Projected Savings (Système *KEFIR*, cf. section 2.3.5)	Matheus et al. [185]
Misclassification costs	Freitas [96]
Vague Feelings (*Fuzzy General Impressions*)	Liu et al. [173]
Anticipation	Roddick et Rice [240]
Interestingness	Shekar et Natarajan [257]

TABLEAU 2.5 : Exemples de mesures d'intérêts subjectives

2.3.4.3 Mesures sémantiques

Selon la classification de Geng et Hamilton [98], les mesures sémantiques se déclinent en deux catégories :

- Les mesures dites d'*utilité* : généralement basées sur des notions de « poids » attribués aux items, aux transactions ou aux clients de manière à mettre en valeur l'utilité selon les contraintes du domaine ;

- Les mesures dites d'*actionnabilité* : généralement orientées métier permettant ainsi d'orienter l'extraction de manière à obtenir des règles pour prendre des décisions.

Le terme « utilité » est utilisé couramment pour définir la « qualité d'être utile » et les utilités sont largement utilisées dans le processus de prise de décision pour exprimer les préférences utilisateurs en fonction des objectifs [97] : *décision = probabilité + utilité*. Selon Shen et al. [258], l'intérêt d'une mesure réside autant dans sa probabilité que dans son utilité. Il présente l'approche *Objective-Oriented utility-based Association (OOA)* permettant d'extraire des règles d'association très corrélées aux objectifs et aux utilités des utilisateurs. De plus, il souligne que son approche permettrait aux décideurs de découvrir les meilleurs stratégies ou opportunités en spécifiant les objectifs maximisant les profits et réduisant les coûts. Dans le milieu médical cela pourrait permettre de trouver le meilleur traitement en répondant au besoin en amont « meilleure efficacité et peu d'effets secondaires ». De nombreuses approches mettant en place des mesures d'utilité ont été proposées et sont récapitulées dans le tableau 2.6.

Mesure sémantique	**Modèle**	**Origine**
Weighted support	*Weights for items*	Support
Normalized weighted support	*Weights for items*	Support
Vertical weighted support	*Weights for transactions*	Support
Mixed weighted support	*Weights for both items and transactions*	Support
OOA	*Weights for target attributes*	Support
Marketshare	*Weight for each transaction*	Confiance
Count support	*Weights for items and cells in dataset*	Support
Amount support	*Weights for items and cells in dataset*	Support
Count confidence	*Weights for items and cells in dataset*	Confiance
Amount confidence	*Weights for items and cells in dataset*	Confiance
Yao et al.'s	*Weights for items and cells in dataset*	Support

TABLEAU 2.6 : Exemples de mesures d'utilité [50]

Récemment, Wang et al. [290] et Lin et al. [167] suggèrent d'ajouter de la valeur aux règles d'association, par exemple en associant une valeur de marge à chaque item de la base de données, permettant ainsi de trier les règles par gain économique. Actuellement, les mesures d'utilité connues sont des extensions du support ou de la confiance (cf. tableau 2.6). De plus, il existe rarement une mesure d'utilité applicable face à une problématique définie. De nombreuses mesures répondent à un besoin bien précis et sont aujourd'hui non généralisables.

2.3.5 Règles actionnables

Matheus et al. [185] présentent l'intérêt de la déviation via le potentiel de bénéfice engendré par une action. Cette approche est implémentée dans le système KEFIR (*KEy FIndings Reporter*) (cf. figure 2.14), un système permettant de découvrir et d'expliquer les principales conclusions des règles d'association, générant lorsque cela est approprié des recommandations aux utilisateurs pour déclencher des actions.

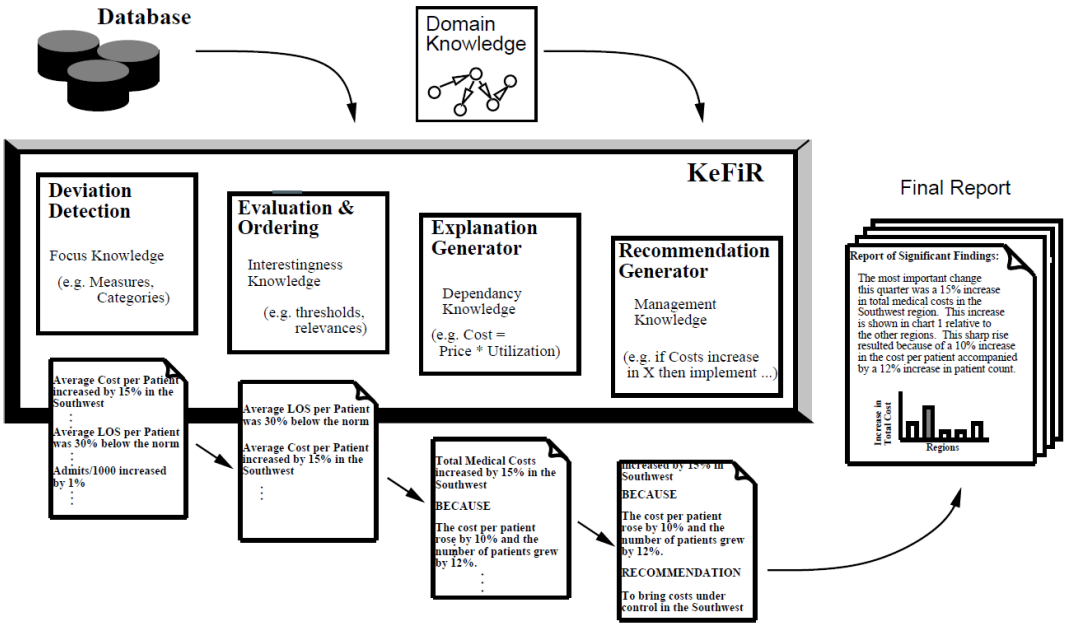


FIGURE 2.14 : Processus du système *Key Findings Reporter* [184]

Ras et al. ont introduit la notion d'*Action Rules* en 2000 [231], approfondie par la suite [230, 279, 290], comme étant une « règle décrivant une transition possible de l'objet d'un état à un autre ». Ras et al. distinguent les objets dits *stables* des *flexibles* pour construire les règles d'actions. Ils définissent les règles d'association d'actions en proposant un algorithme basé sur une contrainte de support. Des actions sont définis par les experts métier et considérées comme des itemsets fréquents. Cependant, en aucun cas cet algorithme ne permet d'élaguer les règles d'association extraites. Ce concept est fortement similaire à la notion d'*Intervention* proposée par Greco et al. en 2005 [107].

Tzacheva et Ras [279] introduisent par la suite les notions de coût (*cost*) et de faisabilité (*feasibility*) d'une règle actionnable, la coût étant une mesure subjective et la faisabilité une mesure objective. L'algorithme ARAS (*Action Rules discovery based on Agglomerative Strategy*) permet d'extraire des règles d'association actionnables [232].

Wickramaratna et al. [296] proposent une approche basée sur la construction de « classes de règles d'association » pour mettre en évidence les co-occurences des items dans une base de transactions. Les auteurs ont mis en place une variante de l'algorithme *ITemset Trees* (*IT-trees*) pour obtenir l'ensemble des règles d'association dont l'antécédent contient au moins un item non acheté pour les clients ciblés. Cette technique est par la suite combinée à d'autres techniques bayésiennes.

Webb [293] présente un ensemble de techniques génériques à base d'extraction de règles d'association permettant de différencier les « fausses » des « vraies » connaissances à l'aide de tests d'hypothèses statistiques arbitraires fournissant un contrôle stricte sur le risque de fausses découvertes. Zhao et al. [310] proposent une nouvelle

approche d'exploration des données afin d'identifier rapidement des règles d'association actionnables. Liu et al. [175] présentent une manière d'identifier des règles non actionnables. La technique proposée fonctionne en deux phases : i) une génération des règles puis un élagage des règles non significatives et ii) un élagage des règles non-actionnables parmi les règles pré-sélectionnées.
Un des défis actuel de l'extraction des règles d'association est de savoir ce que l'utilisateur doit exploiter ou actionner avec de nouvelles règles. Imielinski et Virmani [138] offrent la possibilité d'avoir les mêmes fonctionnalités que dans un DBMS (*DataBase Management System*) pour la découverte de connaissances. Ils introduisent le KDMS (*Knowledge Discovery Management System*) pour faire face au DBMS. L'extraction de connaissances sous forme de règles d'association peut être actionnée de différentes manières. Nous présentons ci-dessous la classification de Imielinski et Virmani [138] de l'actionnabilité des règles d'association :

- ***Typicality*** : sélectionner les individus typiques et atypiques dans une base de données. Par exemple, un individu typique représente celui qui respecte les règles les plus fortes (support et confiance élevés) et l'atypique celui qui viole les règles en question ;
- ***Characteristic of*** : mettre à disponibilité des utilisateurs des cubes de données (de transactions vérifiant une conjonction d'items) à l'aide de règles plus ou moins importantes pour les experts métier ;
- ***Changing patterns*** : mesurer les impacts sur les règles d'association en cas de changement de valeurs pour les individus suite au déclenchement d'une ou de plusieurs actions (une campagne marketing par exemple) ;
- ***Best N*** : trouver les meilleures règles supportées par un certain nombre d'individus. Cette application peut être utile en marketing pour cibler de très bons clients avec des packs produits susceptibles d'être achetés. De plus, le ROI peut être calculé facilement.
- ***Clustering*** : regrouper des individus en fonction de leurs similarités face aux règles d'association respectées ou violées, cela engendrant la connaissance de segments de marché ou de produits intéressants à prospecter ;
- ***What if*** : rechercher la meilleure action à mener pour répondre à un objectif défini par les experts métier. Par exemple, quelle action enclencher pour développer au maximum le potentiel d'un client ?
- ***Knowledge summarization*** : trouver les règles qui caractérisent une population. Par exemple quelle(s) règle(s) caractérise(nt) au mieux une sous population définie à l'aide de critères métier. Cette approche peut être considérée comme une *méta-requête* sur les règles d'association et non directement sur les données ;
- ***Cross verification*** : vérifier que les règles respectant une problématique se retrouvent sur une autre problématique même en renommant certaines propriétés ;

- ***Interesting rules*** : rechercher si une règle peut être qualifiée de potentiellement intéressante en sélectionnant les individus sur lesquels on peut l'appliquer. Si vous trouvez que les individus qui achètent des plaques de plâtres achètent des vis en Vendée mais que tous les individus français achètent des vis, alors l'actionnabilité de la règle n'aura pas d'intérêt pour les experts métier.

Ainsi, nous pouvons définir la notion de *règle d'association actionnable* [311, 312]. Une règle d'association est actionnable « si l'utilisateur peut enclencher une action à son avantage basée sur la règle » [173].

2.4 Positionnement

L'ECD se concentre principalement sur la découverte de modèles satisfaisant des indicateurs. Cependant, les experts métier souhaitent mettre en place des actions en s'appuyant sur ces connaissances. C'est pourquoi, nous concentrons particulièrement nos efforts sur la phase de post-traitement de l'ECD (cf. figure 2.15).

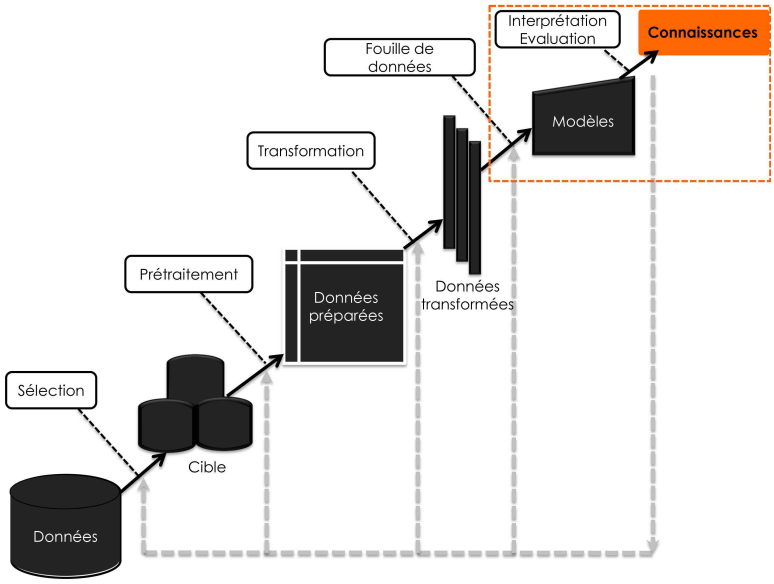


FIGURE 2.15 : Phase de post-traitement de l'ECD [90]

À travers la littérature et les travaux menés, on peut remarquer que l'ECD a franchi plusieurs caps :

- Dans un premier temps, un effort important, et qui se traduit par une abondante littérature, a été consacré à amélioration de l'efficacité des algorithmes afin de lever le verrou du passage à échelle.

- Puis, ce problème d'efficacité résolu, un nombre croissant de travaux ont pu s'intéresser à l'évaluation de la qualité des connaissances produites, grâce aux

méthodes de validation statistiques des modèles sous-jacents, ainsi qu'en développant de nouvelles mesures d'intérêt pour les décideurs.

Cependant, peu de travaux traitent de l'actionnabilité. Pourtant, cette dernière vise à rendre utile à l'action la connaissance issue des modèles. L'actionnabilité demeure un verrou scientifique en ECD mais pourtant cruciale pour l'usage des modèles et nécessite d'être complétée par la mesure de profit (cf. figure 2.16).

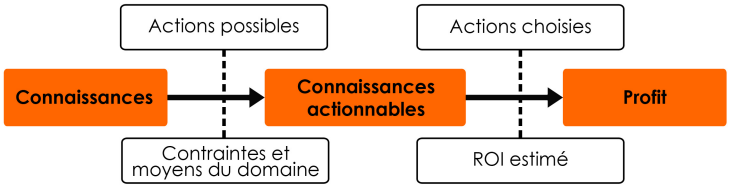


FIGURE 2.16 : Actionnabilité des connaissances issues de l'ECD

La découverte de connaissances actionnables a laissé apparaître des méthodologies pilotées par le domaine. Ces dernières présentent les grands principes à suivre pour extraire des connaissances en prenant en compte les exigences des experts métier. Généralement, ces méthodologies implémentent des techniques de fouille de données guidées par l'action ou par le profit ayant pour finalité principale le CRM. Notamment, la technique de l'extraction des règles d'association est couramment utilisée pour développer la valeur client. En effet, les règles représentent un modèle explicite pour aider les commerciaux à développer la valeur de leurs clients. Des systèmes basés sur des règles dites actionnables et appuyés sur des mesures d'utilité sont apparus mais ont été peu validés et peu appliqués dans des contextes réels.

2.5 Conclusion

En définissant les notions d'*extraction de connaissances à partir des données*, de *gestion de la relation client*, d'*actionnabilité* et de *retour sur investissement*, ce chapitre établit un cadre formel pour l'actionnabilité de la connaissance répondant aux besoins des experts métier pour la prise de décision. Dans ce cadre, nous avons pu présenter le D^3M (*Domain Driven Actionable Knowledge Delivery*) offrant des méthodologies, des techniques et des outils de fouille de données qui s'adaptent aux contraintes des entreprises. Ainsi, nous avons présenté les techniques de fouille de données les plus couramment utilisées pour l'extraction de savoirs actionnables : le *clustering*, les arbres de décision, le *scoring* et les réseaux Bayésiens. Nous nous sommes concentrés davantage sur la technique d'extraction de règles d'association et notamment sur leur actionnabilité. Pour cela nous avons présenté la notion de *règle actionnable*, permettant d'enclencher une action à l'avantage de l'expert métier.
Nous verrons dans le chapitre 3 une présentation des systèmes de recommandation pour fidéliser les clients à travers des suggestions actionnables et profitables, développant ainsi leur valeur économique.

3 Systèmes de recommandation

In this age of information overload, people use a variety of strategies to make choices about what to buy, how to spend their leisure time, and even whom to date [...]

Dietmar Jannach, 2011

SOMMAIRE

L'émergence et le développement du commerce électronique ont conduit au progrès des systèmes de recommandation, un domaine de recherche en plein essor dont les premiers articles fondateurs sont apparus dans le milieu des années 90 [128, 236]. Ces derniers permettent aux entreprises de filtrer l'information, puis de recommander de manière proactive des produits à leurs clients en fonction de leurs préférences. Recommander des produits et des services peut renforcer la relation entre l'acheteur et le vendeur, et donc augmenter les bénéfices [309]. Les systèmes de recommandation doivent veiller à accroître la satisfaction des utilisateurs plutôt que d'émettre des suggestions en rapport avec une politique commerciale. Cette situation peut notamment être rencontrée dans le contexte des applications de commerce électronique où l'hébergeur cherche à influencer les utilisateurs. À titre d'exemple, *Amazon* a déjà reconnu avoir eu recours à de fausses recommandations afin de favoriser les nouveaux articles vestimentaires de leurs partenaires [297].
Ces dernières années sont révélatrices de l'utilisation des systèmes de recommandation sur le Web à travers les films [1], les livres [2] et la musique [3]. Schafer et al. [249] présentent une classification détaillée des systèmes de recommandation pour le commerce électronique, et élucident la façon dont ils peuvent être utilisés pour fournir un service personnalisé fidélisant le client. Actuellement, les moteurs de recommandation reposent sur trois paradigmes [238] :

1. L'un basé sur le contenu (*content-based filtering*) ;
2. L'autre sur la collaboration (*collaborative filtering*) ;
3. Le dernier sur l'hybridation des deux premières approches (*hybrid filtering*).

Dans ce chapitre nous présentons tout d'abord les différentes formes de recommandations existantes en nous concentrant sur les recommandations personnalisées. Ensuite, nous présentons les deux approches les plus communément utilisées, à savoir les systèmes de recommandation collaboratifs (*collaborative filtering*) et thématiques ou basés sur le contenu (*content-based filtering*). Nous nous concentrons sur les techniques utilisées dans le filtrage collaboratif. Enfin, nous présentons les domaines d'applications et les systèmes collaboratifs les plus en vogue aujourd'hui.

3.1 Types de recommandations

Les recommandations peuvent être catégorisées en fonction des objets (*items*) suggérés, des données d'entrée (utilisateurs) et des objectifs métier. Dans le tableau 3.1, quatre types de recommandation de contenus sont présentés faisant face à quatre stratégies métier différentes [227].

1. http://www.netflix.com
2. http://www.amazon.com
3. http://www.last.fm

1. **Éditoriale** (« *post-it* ») permettant de mettre en avant des contenus, pour tous les utilisateurs, sur des critères tels que la nouveauté ou la popularité par exemple ;
2. **Sociale** (« *Facebook* ») permettant à l'utilisateur de noter, commenter, et recommander des contenus à son réseau social et de partager son propre univers avec ses amis ou d'autres utilisateurs du service ;
3. **Contextuelle** (« *more like this* ») suggérant à l'utilisateur des recommandations similaires au contenu qu'il est en train de consommer : par exemple des contenus du même auteur, avec les mêmes acteurs/artistes ou sur le même thème ;
4. **Personnalisée** (« *broaden my horizon* ») visant à mettre en avant des contenus liés aux préférences, au profil et à l'historique de l'utilisateur.

Éditoriale	**Contextuelle**
Top 10, les mieux notées *Nouvelles sorties, la sélection du site* *Les plus consultés en ce moment*	*Du même artiste* *Ceux qui ont aimé... aiment également...* *Les clients ont vu cet article puis ont vu ceux-là*
Sociale	**Personnalisée**
Votre réseau social vous recommande *Votre communauté / vos amis aiment...* *Ce que les auditeurs écoutent en ce moment*	*Susceptible de vous plaire* *Recommandations pour vous* *Bientôt disponible pour vous*

TABLEAU 3.1 : Quatre types de recommandations pour quatre stratégies [227]

Le leader de la musique en ligne, *Deezer* [4], souhaitant avant tout faire découvrir son catalogue à ses utilisateurs privilégie les recommandations éditoriale et contextuelle. A contrario, *Pandora* [5], souhaite davantage développer son audience et utilise les recommandations sociale et personnalisée pour fidéliser les utilisateurs en leur permettant de créer leur propre univers.
Dans les sous sections suivantes, nous détaillons l'intérêt et les objectifs stratégiques des quatre types de recommandations en nous inspirant des travaux de Poirier [227] et Jannach et al. [141].

3.1.1 Recommandation éditoriale

La recommandation dite *éditoriale* est utilisée lorsqu'aucune donnée d'entrée n'est disponible. Ce type de recommandation a pour objectif de persuader l'utilisateur de s'intéresser à un ou plusieurs items et ainsi provoquer l'acte d'achat. Les recommandations ne sont pas personnalisées aux utilisateurs mais qualifiées de « génériques ». Par exemple, de nombreux sites Internet mettent en avant des nouveautés

4. http://www.deezer.com
5. http://www.pandora.fm

(cf. figure 3.1), des articles fréquemment achetés (cf. figure 3.2), des promotions (cf. figure 3.3) ou simplement des items à forte valeur ajoutée que l'entreprise souhaite destocker à l'aide de ventes *flash* (cf. figure 3.4).

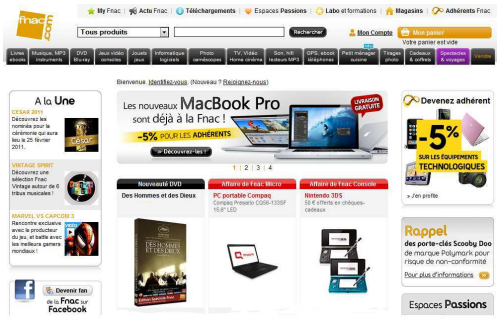


FIGURE 3.1 : Nouveautés *A La Une* sur le site de la *Fnac*

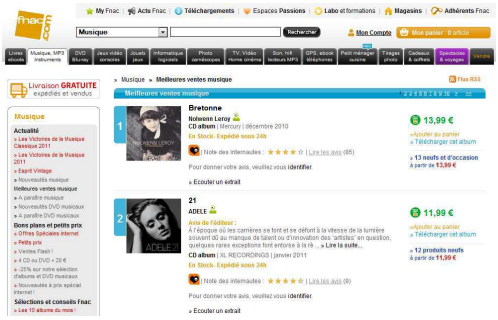


FIGURE 3.2 : Meilleures ventes sur le site de la *Fnac*

FIGURE 3.3 : Promotions du jour sur le site de *WalMart*

FIGURE 3.4 : Ventes *flash* sur le site de *CDiscount*

3.1.2 Recommandation sociale

La recommandation dite *sociale* est réalisée par des utilisateurs différents de l'utilisateur courant. Par exemple, il peut s'agir d'internautes navigant sur *YouTube* ou *Flixter*, ou de consommateurs d'*Amazon* (cf. figure 3.5) ou de *Priceminister*, recommandant à l'intérieur même de leur communauté en transmettant leurs appréciations. D'autres sites proposent également de renseigner les « coups de cœur » (cf. figure 3.6), ces derniers permettant d'enrichir des listes de suggestions des communautés d'internautes.

Le principe du réseau social *Myspace* [6] est légèrement différent. Les artistes ou producteurs de contenus recommandent d'autres artistes dont ils apprécient les œuvres, offrant ainsi l'opportunité à chacun d'être consulté.

6. http://www.myspace.com

FIGURE 3.5 : Exemple de recommandation sociale sur *Amazon*

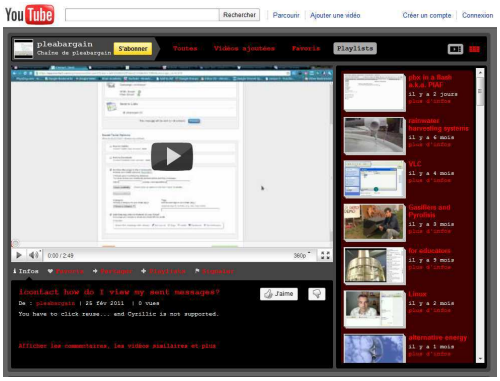


FIGURE 3.6 : *PlayList* conseillée sur *YouTube*

3.1.3 Recommandation contextuelle

La recommandation dite *contextuelle* consiste à proposer des items « proches » de l'item consulté. Une première démarche rapproche les items dont les caractéristiques descriptives sont similaires. Ensuite, une approche plus complexe, telle que celle d'*Amazon*, recommande une liste d'items appréciés par d'autres utilisateurs ayant également apprécié l'item consulté (cf. figure 3.9). *Flickr*[7] ou *IMDb*[8] utilisent les étiquettes (ou *tags*) pour rapprocher les items. *Pandora* rapproche les morceaux musicaux de son catalogue en analysant plus de quatre cents caractéristiques sonores, déterminant ainsi des « profils musicaux ». Enfin, *Genius*, le système de recommandation d'*Apple* permet d'obtenir des suggestions d'applications directement sur son *iPhone* (cf. figure 3.7) en supprimant à sa propre initiative les mauvaises recommandations, ou de se faire recommander l'achat de musiques ou de films sur l'*Itunes Store* (cf. figure 3.8).

3.1.4 Recommandation personnalisée

La recommandation dite *personnalisée* a pour objectif de recommander à l'utilisateur les contenus ou services les plus susceptibles de l'intéresser. Ces recommandations sont établies en fonction du comportement d'achat ou de navigation de l'utilisateur courant. Les enjeux de la recommandation personnalisée sont nombreux et bénéficient autant à l'utilisateur qu'au fournisseur de services. Tout d'abord, le système de recommandation peut être considéré comme « remplaçant du vendeur », qui généralement capte et capitalise sur les préférences des utilisateurs afin de les guider dans leurs choix. Ensuite, l'utilisateur lui-même est avantagé par un gain de temps et une découverte d'items souvent cachés auxquels il n'aurait pas pensé. Enfin, du côté fournisseur de la recommandation, satisfaction et fidélisation clients viennent souvent engendrer des bénéfices et guider la stratégie marketing de l'entreprise.

7. http://www.flickr.com
8. http://www.imdb.com

FIGURE 3.7 : Recommandation d'applications *Iphone*

FIGURE 3.8 : Recommandation d'achats de musique sur *Itunes*

Par exemple, le site *Allociné* propose à ses utilisateurs des suggestions de films (cf. figure 3.10) encore non visualisés en fonction des préférences utilisateurs et des notes attribuées aux films déjà visualisés.

FIGURE 3.9 : Recommandation contextuelle sur *Amazon*

FIGURE 3.10 : Recommandation personnalisée *AlloCiné*

Actuellement, la recommandation personnalisée représente le domaine le plus attractif pour l'utilisateur et le plus intéressant économiquement pour le fournisseur de recommandations [272].

3.2 Systèmes collaboratifs

3.2.1 Historique

Le terme *collaborative filtering* a été proposé par David Golberg et ses collaborateurs chez Xerox en 1992 avec la mise en place du système de recommandation personnalisées *Tapestry* [100]. Deux ans plus tard, en 1994, Paul Resnick du MIT (*Massachusetts Institute of Technology*) et ses collaborateurs de l'université du Minnesota ont proposé l'architecture *GroupLens* pour recommander des articles dans les *newsgroup* [236]. La librairie *Amazon* et plus particulièrement Greg Linden a popularisé le filtrage collaboratif avec sa fonction « les utilisateurs qui ont aimé ce livre ont aussi aimé tel autre livre ». En 1998, Brin et Page ont publié leur algorithme *PageRank* et lancé *Google*. La même année chez *Microsoft*, John S. Breese et ses collaborateurs ont publié un article charnière, *Empirical Analysis of Predictive Algorithms for Collaborative Filtering* [37] dans lequel figure une comparaison détaillée des divers algorithmes de filtrage collaboratif.
Dans les années 2000, les algorithmes de filtrage collaboratif étaient basés sur les réseaux bayésiens ou les réseaux de neurones avec une approche basée sur l'utilisateur. En 2001, *Amazon* dépose un brevet introduisant le filtrage collaboratif basé sur l'item [169] ; ce type d'algorithme étant également publié la même année et de façon indépendante par la communauté *GroupLens*. En 2006, la compagnie *Netflix* annonce son challenge avec une récompense très attrayante, rendant ainsi disponible un ensemble de données réelles et volumineuses pour évaluer les systèmes de recommandation.

3.2.2 Terminologie

Une définition formelle de la recommandation a été introduite par Adomavicius et Tuzhilin [2] :

DÉFINITION 5
*Soit C l'ensemble de tous les utilisateurs et P l'ensemble de tous les items qui peuvent être recommandés. Soit U un ensemble ordonné et $u : C \times P \rightarrow U$ une fonction mesurant l'**intérêt** ou l'**utilité** porté par l'utilisateur $c \in C$ à l'item $p \in P$. Dès lors, pour chaque utilisateur $c \in C$, le système de recommandation sélectionne l'item $p' \in P$ qui maximise l'intérêt ou l'utilité de c.*

$$\forall c \in C, \quad p'_c = arg \max_{p \in P} u(p, c) \tag{3.1}$$

L'intérêt ou l'utilité d'un utilisateur c pour un item p, noté $u(p, c)$ est généralement représenté par une note ou une mesure économique telle que le chiffre d'affaires ou la marge nette. Afin de prédire cet intérêt, des connaissances sur l'utilisateur ou sur l'item doivent être assimilées par le système de recommandation. De plus, les mesures déjà portées sur le contenu par certains utilisateurs peuvent également être collectées. Ces informations sont regroupées dans une matrice appelée « matrice d'usage ». Un exemple de matrice d'usage binaire (*aime/n'aime pas*) est présentée dans le tableau 3.2. Par exemple, l'utilisateur c_2 apprécie l'item p_3 mais pas l'item p_2.

En revanche, l'utilisateur c_5 apprécie l'ensemble des items excepté l'item p_5. Enfin, l'item p_5 n'est apprécié par aucun utilisateur ayant donné son avis.

	p_1	p_2	p_3	p_4	p_5
c_1	☺			☺	
c_2		☹	☺		
c_3		☺			
c_4			☹	☺	☹
c_5	☺	☺	☺	☺	☹

TABLEAU 3.2 : Matrice d'usage binaire (*aime/n'aime pas*) de cinq utilisateurs et cinq items

Ces informations peuvent également cacher une valeur économique, c'est-à-dire « a acheté pour telle somme ou n'a pas acheté » (cf. tableau 3.3) ou alors une note sur une échelle prédéfinie « a adoré, a apprécié, a détesté, etc. ». Elles peuvent également se mesurer sur un nombre plus élevé de classes : « a voté 1, 2, 3, 4 ou 5 étoiles » (cf. tableau 3.4). L'objectif du système de recommandation est de prédire les mesures d'intérêts *utilisateur-item* manquantes. En d'autres termes, remplir les cases vides $u(p, c)$ de la matrice d'usage en évaluant si l'item p intéresse l'utilisateur c.

	p_1	p_2	p_3	p_4	p_5
c_1	2500 €			3600 €	
c_2			1900 €		
c_3		850 €			
c_4				2400 €	
c_5	8000 €	7000 €	1200 €	2100 €	

TABLEAU 3.3 : Matrice d'usage de chiffres d'affaires

	p_1	p_2	p_3	p_4	p_5
c_1	4			5	
c_2		1	4		
c_3		3			
c_4			2	4	1
c_5	5	5	3	4	1

TABLEAU 3.4 : Matrice d'usage de votes

Trois types d'approches sont principalement utilisées [201, 227] : le filtrage basé sur le contenu, le filtrage collaboratif et le filtrage hybride.

1. Le filtrage **basé sur le contenu** repose sur les variables descriptives des items afin de les comparer et ou de les corréler au profil des utilisateurs [216]. Chaque utilisateur du système possède un profil le décrivant à travers ses centres d'intérêts par exemple. Lors de l'arrivée d'un nouvel item, le système compare la représentation de l'item avec le profil utilisateur afin de prédire la mesure d'intérêt que pourrait porter l'utilisateur sur l'item. Les items sont alors recommandés en fonction de leur proximité aux utilisateurs.

2. Le filtrage **collaboratif** repose sur les mesures d'intérêts utilisateurs/items stockées dans la matrice d'usage. Deux approches de filtrage collaboratif se distinguent. L'approche basée sur les utilisateurs [236] consistant à comparer les utilisateurs entre eux et à retrouver ceux ayant des mesures proches, les prédictions étant ensuite réalisées par voisinage. L'approche basée sur les items [245]

consistant à rapprocher les items appréciés ou non par des utilisateurs similaires et à prédire les mesures d'intérêts des utilisateurs en fonction des items les plus proches de ceux qu'ils ont déjà mesuré.

3. Le filtrage **hybride** exploite aussi bien l'approche collaborative que thématique. Le système hybride peut faire appel à des sources d'informations complémentaires telles que des données démographiques ou sociales [215]. Différentes méthodes d'hybridation peuvent être envisagées afin de combiner les modèles. Par exemple, il est possible d'appliquer séparément le filtrage collaboratif et d'autres techniques de filtrage pour générer des recommandations dites « candidates », et combiner ces ensembles de recommandations par *pondération*, *cascade* ou encore *bascule* [49, 227].

Dans la section suivante, nous nous intéressons plus formellement aux deux approches les plus communément utilisées : les filtrages collaboratif et thématique.

3.2.3 Filtrage collaboratif

Goldberg et al. [100] ont défini le filtrage collaboratif (ou *Collaborative filtering*) comme « une technique employant les comportements connus d'une population pour prévoir les futurs agissements d'un individu à partir de l'observation de son attitude dans un contexte donné ». Sarwar et al. [245] désignent le filtrage collaboratif comme « les systèmes de recommandation se basant sur les opinions et évaluations d'un groupe d'utilisateurs afin de suggérer un ou plusieurs items ». Le principe général consiste lorsqu'un nouvel utilisateur se présente, à rechercher par comparaison les utilisateurs qui, par leurs préférences connues, semblent avoir des comportements similaires [153, 236]. Cette approche utilise essentiellement les données contenues dans la matrice d'usage. La matrice peut être construite en « sondant » les comportements des utilisateurs ou en proposant à ces derniers de déclarer eux-mêmes leurs mesures d'intérêts sur les items. Par conséquent, deux types de filtrage collaboratif ont été définis [141] :

- Le filtrage collaboratif **passif** reposant sur l'analyse des comportements des utilisateurs : par exemple, les achats réalisés dans un magasin ou les pages web visitées sur une période de temps prédéfinie [80]. Aucune information n'est demandée aux utilisateurs. Dès lors, les données récupérées ne contiennent pas de biais dit de déclaration ;
- Le filtrage collaboratif **actif** reposant sur des données déclarées par les utilisateurs telles que les notes des films visualisés sur *NetFlix* [153].

La matrice d'usage peut être exploitée de deux façons différentes [141] :

1. L'orientation utilisateurs (*user-based*) : consistant à calculer les similarités entre utilisateurs à l'aide de leurs profils ;
2. L'orientation items (*item-based*) : consistant à calculer les similarités entre items selon les mesures attribuées par les utilisateurs.

L'objectif étant de prévoir les cases vides de la matrice d'usage, deux approches se distinguent [37] :

- Les **approches basées sur un modèle** (*model-based*), construisant un modèle de prédiction souvent probabiliste sur une partie de la base de données ;
- Les **approches basées sur la mémoire** (*memory-based*), comparant, pour chaque recommandation, l'utilisateur courant à l'ensemble de la base de données.

Les approches basées sur un modèle mettent en œuvre des méthodes issues de l'apprentissage automatique comme les modèles bayésien ou les méthodes de *clustering*. Ces méthodes sont généralement performantes mais ont un coût de construction et de fonctionnement plus important que les méthodes basées sur la mémoire [51, 265]. Néanmoins, dans le cas de données dispersées, ces méthodes semblent plus efficaces que les approches basées sur la mémoire. Pour le lecteur intéressé, une description précise des approches basées sur les modèles est proposée par Su et Khoshgoftaar [265].
Les approches basées sur la mémoire estiment les similarités entre les lignes (*user-based*) ou les colonnes (*item-based*) de la matrice d'usage [248]. Dans le cas des utilisateurs, il s'agit de rapprocher les utilisateurs ayant les mêmes comportements en fonction de leurs mesures d'intérêts. Dans le cas des items, on cherche à rapprocher les items ayant été mesurés de la même façon par les mêmes utilisateurs. Ce type de recommandation est donc réalisé en deux étapes : i) calculer les similarités entre les utilisateurs ou les items de la matrice, et ii) compléter les cases vides de la matrice à l'aide d'une fonction de prédiction. Enfin, le système prédit les mesures que l'utilisateur aurait attribué aux items et suggère les N premiers items, appelés *Top-N* recommandations.
Afin de présenter les différentes approches possibles pour chacune des deux étapes du filtrage collaboratif, introduisons les notations suivantes :

- Un ensemble C de n utilisateurs $\{c_1, c_2, \dots c_n\}$;
- Un ensemble P de m items $\{p_1, p_2, \dots p_m\}$;
- Un ensemble U de mesures $u(p, c)$ de l'utilisateur $c \in C$ pour l'item $p \in P$;
- $E_c \subseteq P$ l'ensemble des items mesurés par l'utilisateur c ;
- $E_p \subseteq C$ l'ensemble des utilisateurs ayant mesuré l'item p.

3.2.3.1 Calcul des similarités

Le calcul des similarités consiste à mesurer la similitude entre les lignes ou les colonnes de la matrice d'usage. Le choix de la mesure utilisée dépend généralement de la nature des vecteurs. Par exemple, si les vecteurs contiennent uniquement des données binaires (cf. figure 3.2), la distance de *Jaccard* peut être utilisée. Considérons deux vecteurs a et b, la mesure de similarité S utilisant la distance de *Jaccard* peut être définie comme suit [238] :

$$S_{Jaccard}(a,b) = \frac{|E_a \cap E_b|}{|E_a \cup E_b|} \tag{3.2}$$

Il s'agit du rapport entre la cardinalité de l'intersection des ensembles considérés et la cardinalité de l'union des ensembles. En revanche, si les données sont des notes (cf. figure 3.4) ou des valeurs de chiffres d'achats (cf. figure 3.3), les deux mesures les plus couramment utilisées sont la *similarité Cosinus* et la *similarité de Pearson* définies comme suit pour deux produits a et b [238] (et symétriquement dans le cas de deux items a et b) :

$$S_{Cosinus}(a,b) = \frac{\sum_{\{x \in E_a \cap E_b\}} u(a,x) \times u(b,x)}{\sqrt{\sum_{\{x \in E_a \cap E_b\}} u(a,x)^2 \sum_{\{x \in E_a \cap E_b\}} u(b,x)^2}} \tag{3.3}$$

$$S_{Pearson}(a,b) = \frac{\sum_{\{x \in E_a \cap E_b\}} (u(a,x) - \overline{u_a}) \times (u(b,x) - \overline{u_b})}{\sqrt{\sum_{\{x \in E_a \cap E_b\}} (u(a,x) - \overline{u_a})^2 \sum_{\{x \in E_a \cap E_b\}} (u(b,x) - \overline{u_b})^2}} \tag{3.4}$$

où $\overline{u_a}$ (respectivement $\overline{u_b}$) représente la moyenne des valeurs contenus dans le vecteur a (respectivement b). Une matrice de similarités *Items-Items* (cf. tableau 3.5) ou *Utilisateurs-Utilisateurs* (cf. tableau 3.6) est alors établie et utilisée pour prédire les valeurs manquantes, i.e les non votes ou les non achats.

	p_1	p_2	p_3	p_4	p_5
p_1		$S(p_1,p_2)$	$S(p_1,p_3)$	$S(p_1,p_4)$	$S(p_1,p_5)$
p_2	$S(p_2,p_1)$		$S(p_2,p_3)$	$S(p_2,p_4)$	$S(p_2,p_5)$
p_3	$S(p_3,p_1)$	$S(p_3,p_2)$		$S(p_3,p_4)$	$S(p_3,p_5)$
p_4	$S(p_4,p_1)$	$S(p_4,p_2)$	$S(p_4,p_3)$		$S(p_4,p_5)$
p_5	$S(p_5,p_1)$	$S(p_5,p_2)$	$S(p_5,p_3)$	$S(p_5,p_4)$	

TABLEAU 3.5 : Matrice de similarités *Items-Items*

	c_1	c_2	c_3	c_4	c_5
c_1		$S(c_1,c_2)$	$S(c_1,c_3)$	$S(c_1,c_4)$	$S(c_1,c_5)$
c_2	$S(c_2,c_1)$		$S(c_2,c_3)$	$S(c_2,c_4)$	$S(c_2,c_5)$
c_3	$S(c_3,c_1)$	$S(c_3,c_2)$		$S(c_3,c_4)$	$S(c_3,c_5)$
c_4	$S(c_4,c_1)$	$S(c_4,c_2)$	$S(c_4,c_3)$		$S(c_4,c_5)$
c_5	$S(c_5,c_1)$	$S(c_5,c_2)$	$S(c_5,c_3)$	$S(c_5,c_4)$	

TABLEAU 3.6 : Matrice de similarités *Utilisateurs-Utilisateurs*

3.2.3.2 Prédiction

La prédiction des valeurs consiste à calculer l'intérêt qu'un utilisateur pourrait porter sur un ou plusieurs items encore non mesurés. Plus formellement, il s'agit de remplir les cases vides de la matrice d'usage à l'aide de la matrice de similarités *Utilisateurs-Utilisateurs* ou *Items-Items*.

Prédire à l'aide de la matrice *Utilisateurs-Utilisateurs*

Le principe consiste à rechercher les utilisateurs possédant le même comportement que l'utilisateur courant [236]. Dès lors, les recommandations sont prédites en fonction des mesures des utilisateurs proches. Soit, $S(c_1, c_2)$ la fonction de similarité entre deux utilisateurs $c_1 \in C$ et $c_2 \in C$. La mesure de l'utilisateur c_1 pour l'item p, notée $u(p, c_1)$, est la somme des mesures déjà faites sur p pondérées par les similarités avec c_1.

$$u(p, c_1) = \frac{\sum_{\{c_2 \in C | p \in E_{c_2}\}} S(c_1, c_2) \times u(p, c_2)}{\sum_{\{c_2 \in C | p \in E_{c_2}\}} S(c_1, c_2)} \tag{3.5}$$

Néanmoins, un problème majeur du filtrage collaboratif est la notation des utilisateurs. En effet, un utilisateur peut, s'il considère que la perfection n'existe pas, ne jamais affecter la note maximale à un item et donc répartir ses notes de 1 à 4 (si les notes possibles vont de 1 à 5). À l'inverse, un utilisateur différent peut, s'il n'aime pas noter trop sévèrement, répartir les notes qu'il attribue de 2 à 5. Pour pallier ce problème, la moyenne des notes de l'utilisateur c_1 est introduite :

$$u(p, c_1) = \overline{u_{c_1}} + \frac{\sum_{\{c_2 \in C | p \in E_{c_2}\}} S(c_1, c_2) \times (u(p, c_2) - \overline{u_{c_2}})}{\sum_{\{c_2 \in C | p \in E_{c_2}\}} S(c_1, c_2)} \tag{3.6}$$

où $\overline{u_{c_1}}$ (respectivement $\overline{u_{c_2}}$) représente la moyenne des notes de l'utilisateur c_1 (respectivement c_2).

Prédire à l'aide de la matrice *Items-Items*

L'intérêt pour les approches basées sur les items est plus récent que celui pour les approches basées sur les utilisateurs [84, 245]. Le site Internet *Amazon* [169] a mis en avant cette approche avec un système construisant une matrice de relation entre les items en se basant sur la base de données d'achats des utilisateurs. Comme pour les approches basées sur les utilisateurs, les performances varient en fonction de la mesure de similarité utilisée et en fonction du nombre d'items proches considérés [65]. Une première façon de calculer la note d'un utilisateur c_1 sur un item p_1 ne prend pas en compte les moyennes de notes.

$$u(p_1, c_1) = \frac{\sum_{\{p_2 \in E_{c_1}\}} S(p_1, p_2) \times u(p_2, c_1)}{\sum_{\{p_2 \in E_{c_1}\}} S(p_1, p_2)} \tag{3.7}$$

Pour pallier les différences d'utilisations des mesures, la moyenne des notes de chaque utilisateur est introduite :

$$u(p_1, c_1) = \overline{u_{p_1}} + \frac{\sum_{\{p_2 \in E_{c_1}\}} S(p_1, p_2) \times (u(p_2, c_1) - \overline{u_{p_2}})}{\sum_{\{p_2 \in E_{c_1}\}} S(p_1, p_2)} \tag{3.8}$$

où $\overline{u_{p_1}}$ (respectivement $\overline{u_{p_2}}$) représente la moyenne des mesure reçues par l'item p_1 (respectivement p_2).

Le bon fonctionnement du filtrage collaboratif nécessite une quantité importante de mesures et d'items « durables », c'est-à-dire des items qui ont un cycle de vie conséquent pour que les utilisateurs aient le temps de les mesurer et que le système puisse les recommander à temps. Concrètement, le filtrage collaboratif n'est pas forcément adapté à un site Internet d'actualités en ligne. Une autre approche, non basée sur le comportement des utilisateurs peut être utilisée afin de pallier le manque de données. Il s'agit du filtrage thématique que nous abordons dans la section suivante.

3.2.4 Filtrage thématique

Le filtrage thématique (ou *Content-based filtering*) [188, 199, 216] consiste à établir des recommandations à l'aide d'« attributs ». Ces derniers, parfois appelés « descripteurs », « caractéristiques descriptives » ou « variables descriptives » caractérisent les items. Plus formellement, les items sont représentés par un vecteur $X = \{x_1, x_2, \ldots x_n\}$ de n composantes. Chaque composante représente un attribut et peut contenir des valeurs binaires, numériques ou catégoriques. Par exemple, pour recommander des films, les attributs peuvent être le genre, le réalisateur, l'année de production, le nombre de récompenses, etc. À partir des variables descriptives, le système de recommandation évalue les similarités existantes. Ce type de système est généralement utilisé dans deux cas particuliers : i) pour remplacer le filtrage collaboratif lorsque le manque d'utilisateurs pénalise le système de recommandation, et ii) pour des systèmes de recommandation d'items à court cycle de vie (les actualités ou les séries télévisées par exemple).

3.2.4.1 Variables descriptives

Le filtrage thématique se base sur les variables descriptives des utilisateurs ou sur les caractéristiques descriptives des items.

Variables descriptives des utilisateurs
Parallèlement au filtrage collaboratif (cf. section 3.2.3), les variables descriptives des utilisateurs peuvent être collectées de deux manières différentes :

- *Passive* : considérant l'historique des items notés ou achetés par les utilisateurs ;
- *Active* : proposant aux utilisateurs un questionnaire et leur offrant la possibilité de mesurer les items en fonction de leurs préférences.

En fonction des données collectées, les variables descriptives peuvent contenir des préférences vis-à-vis d'items ou des caractéristiques descriptives sur les utilisateurs ou les items. Dès lors, le système peut recommander des items proches de ceux appréciés par l'utilisateur courant ou filtrer les items proches de ceux dépréciés.

Caractéristiques descriptives des items
Les variables descriptives des items peuvent être représentées de différentes manières. Tout d'abord, en se basant sur les données descriptives des items telles que le

genre d'un film. Lorsque ces données ne sont pas disponibles ou peu informatives, on peut alors analyser l'item et en extraire les *métadonnées*. Il s'agit de la méthode utilisée sur le site de *Pandora*. Les métadonnées constituent le profil de chaque morceaux et des similarités peuvent être ainsi calculées entre les morceaux appréciés par l'utilisateur courant et les morceaux qu'il n'a pas encore écoutés.

Ensuite, les variables descriptives des items peuvent être renseignées par les utilisateurs sous forme textuelle. Il peut s'agir par exemple de données structurées (systèmes des *tags*), mais également de données non structurées telles que des textes descriptifs (*synopsis* par exemple), des critiques journalistiques ou encore des commentaires sur des forums en ligne. Les mots sont souvent pris comme descripteur et subissent généralement des traitements linguistiques comme la lemmatisation [9] par exemple. Une fois les variables descriptives des utilisateurs ou items collectées, des mesures de similarités peuvent être mises en place.

3.2.4.2 Calcul des similarités

Le calcul des similarités dans le cadre du filtrage thématique consiste à calculer les similarités entre les variables descriptives des items ou des utilisateurs.

Les similarités entre variables descriptives peuvent être calculées avec les distances de similarités traditionnelles telles que *Jaccard*, *Cosinus* ou *Pearson* (cf. section 3.2.3.1) ou la méthode des plus proches voisins. Cette tâche peut également être vue comme une tâche de classification. Des outils d'apprentissage automatique sont alors utilisés, comme des classifieurs naïfs bayésiens ou des systèmes à base de règles d'association. Néanmoins, les distances de similarités sont davantage utilisées et semblent offrir de meilleurs résultats [227].

Différentes techniques plus récentes ont également été proposées. La *distance normalisée de Google* [74] calcule le nombre de co-occurences de termes textuels dans les pages répertoriées par le moteur. Deux termes ayant un fort taux d'apparition sur des pages communes sont considérés comme proches. Par exemple, il est censé être plus difficile de produire l'occurrence d'un homme qui en gifle un autre dans la rue qu'un événement dans lequel l'homme demande simplement l'heure. Une application directe de la distance normalisée de *Google* donne une distance entre « rue » et « gifler » qui est sensiblement plus élevée que la distance entre « rue » et « demander ».

Une autre approche comparant les attributs est l'utilisation de graphes. Bothorel et Bouklit [35] proposent une technique de *clustering* de graphes afin de détecter des communautés de nœuds (les attributs) à l'aide des arêtes (une arête reliant deux attributs présents sur un même item) afin de déterminer les similarités entre attributs.

9. La lemmatisation désigne l'analyse lexicale du contenu d'un texte regroupant les mots d'une même famille. Chacun des mots d'un contenu se trouve ainsi réduit en une entité appelée lemme.

3.2.5 Problèmes rencontrés

Malgré leur popularité croissante, les systèmes de recommandation ont subi quelques ratés. L'exemple le plus révélateur est l'article anecdotique du *Wall Street* Journal intitulé « *If TiVo Thinks You Are Gay, Here's How to Set It Straight* ». Cet article décrit la frustration des utilisateurs résultant des mauvaises suggestions fournies par le système de recommandation d'émissions de télévision *TiVo*[10] [272]. Nous présentons dans cette section les problèmes les plus courants dont souffrent les systèmes de recommandation [141, 227, 238].

3.2.5.1 Démarrage à froid

Le démarrage à froid (ou *Cold Start*) se produit lorsque le système ne possède pas assez de données d'usage. Dès lors les performances de prédictions sont détériorées [2]. Dans le cas du filtrage collaboratif, il n'existe aujourd'hui aucune solution à ce problème. Un système ne possédant pas d'utilisateurs ne pourra pas émettre de recommandations. En revanche, si des données « disponibles à froid » peuvent être collectées, alors, le filtrage thématique peut être considéré. Actuellement, il existe trois types de démarrage à froid [48] : le système débutant, le nouvel utilisateur et le nouvel item.

Le système débutant [191] survient lors du démarrage du système : ce dernier ne possède aucune donnée d'entrée, pas plus sur les utilisateurs que sur les items. Les méthodes de filtrage collaboratif ne peuvent fonctionner que s'il existe des informations dans la matrice d'usage. La solution consiste soit à trouver des variables descriptives des items afin d'organiser les items entre eux et inciter les utilisateurs à les parcourir, remplissant ainsi la matrice d'usage, soit à collecter des données externes en fonction du domaine applicatif [192].

Le nouvel utilisateur survient lorsqu'un nouvel utilisateur entre dans le système et qu'aucune donnée sur l'utilisateur n'a été collectée. Plusieurs solutions existent : lui soumettre un questionnaire sur les items (collection de données *active*), ou faire de la recommandation éditoriale (cf. section 3.1.1) afin de l'inciter à parcourir les items et ainsi enrichir le système. Pour éviter cette tâche fastidieuse pour l'utilisateur, certains auteurs proposent d'associer le nouvel utilisateur à un « stéréotype » en exploitant par exemple une source d'informations démographiques externe comme les pages web personnelles des internautes [215].

Le nouvel item survient lorsqu'un nouvel item est inséré dans le système. Dans le cas du filtrage thématique, il s'agit de trouver des variables descriptives des items afin de comparer l'item aux items existants. Dans le cas du filtrage collaboratif, un

10. http://www.tivo.com/

item n'ayant reçu aucune note ou n'ayant jamais été acheté ne peut être recommandé. Il s'agit alors de le rendre visible aux utilisateurs afin d'obtenir un certain nombre de mesures d'intérêt (typiquement le cas des « fausses » recommandations [297]). Ce problème peut également être traité à l'aide d'une approche hybride utilisant par exemple les similarités entre documents [250] ou introduisant des agents intelligents évaluant automatiquement les documents [102]. Des travaux récents [222] soulignent que les métadonnées sont quasiment inutiles pour les systèmes de recommandation, y compris en situation de démarrage à froid ; les données de *logs* de notations étant beaucoup plus informatives.

3.2.5.2 Effet entonnoir

Lors de la mise en place d'un filtrage thématique, les utilisateurs sont indépendants les uns des autres. Ainsi, l'utilisateur courant peut recevoir des recommandations même s'il est seul dans le système. En revanche, cette technique thématique est soumise à l'effet « entonnoir ». En effet, le profil évolue naturellement par restriction progressive sur les thèmes recherchés. Ainsi, l'utilisateur ne reçoit que les recommandations relatives à ses préférences, une fois son profil devenu stable. Par conséquent, il ne peut pas découvrir de nouveaux domaines pouvant potentiellement l'intéresser. À l'inverse le filtrage collaboratif n'est pas soumis à l'effet entonnoir car les utilisateurs peuvent tirer profit des mesures d'intérêt des autres utilisateurs et recevoir des recommandations pour lesquelles les utilisateurs les plus proches ont émis un intérêt.

3.2.5.3 Longue traîne

La longue traîne [7] est un phénomène connu des systèmes de recommandation et plus généralement des statistiques. Il concerne tous les items *non populaires* ou les *nouveaux* items souvent ignorés des systèmes de recommandation collaboratifs. En effet, ces items étant minoritairement mesurés par les utilisateurs, les algorithmes de filtrage ne les considèrent pas ou très peu. Ce phénomène a tendance à s'accentuer lors de l'évolution du système : « certains items sont de plus en plus à la traîne » et il s'avère que la quantité d'items non populaire est souvent beaucoup plus importante que les items recommandés (cf. figure 3.11). Cette problématique est souvent liée à un manque de données des items non populaires. Dès lors, nous pourrions imaginer d'acquérir des données externes pour pallier ce manque de connaissance [211].

3.2.5.4 Principe d'induction

Les systèmes de recommandation se basent sur le principe qu'un utilisateur tend à reproduire son comportement dans le temps. Cependant, ce principe n'est pas nécessairement vrai dans un contexte réel. En effet, un utilisateur peut changer complètement de domaine d'intérêt ou en avoir plusieurs. Pour faire face à ce problème, des techniques de dérive d'intérêt (*Interest drift*) ou de changement de contexte (*Context shifts*) ont vu le jour [24].

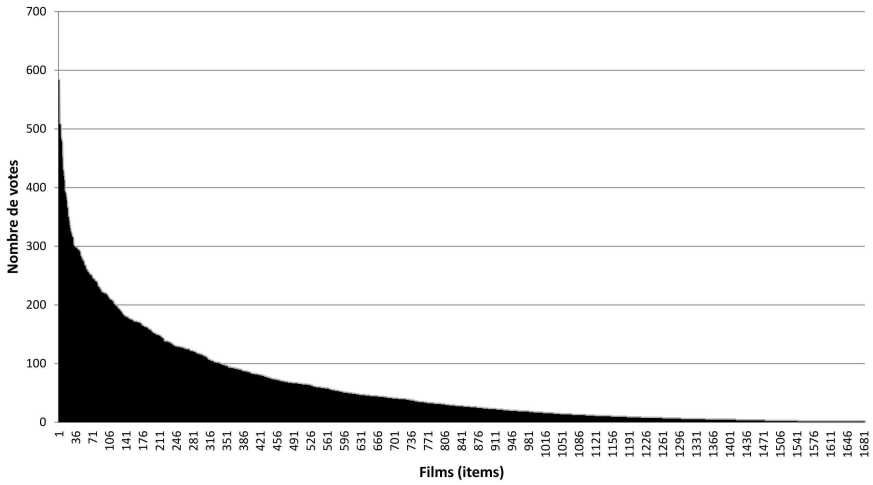


FIGURE 3.11 : Le phénomène de la longue traîne sur les données *MovieLens* [211]

Par exemple, dans le domaine des matériaux de construction, alors qu'il y a quelques années, les clients cherchaient à optimiser le rapport coût/surface de leur logement, ils cherchent maintenant à optimiser la performance énergétique de leur habitat. Dès lors, les mesures d'intérêts face aux matériaux de construction subissent des dérives en fonction des normes imposées.

3.2.6 Évaluation des systèmes de recommandation

L'évaluation des systèmes de recommandation procède généralement par validation croisée [238]. Les données disponibles sont découpées en sous-ensembles : un pour l'apprentissage et l'autre pour la mesure de performance. Ce procédé est ainsi répété plusieurs fois. En pratique, la démarche est la suivante :

1. Choisir de manière aléatoire un certain nombre d'utilisateurs puis leur appliquer l'algorithme d'apprentissage. Les autres utilisateurs constituent l'ensemble de validation ;

2. Pour chaque utilisateur de l'ensemble de validation, fournir une sous partie des mesures. Par exemple, blanchir aléatoirement 20 % des mesures. Ainsi, on cherche à prédire les notes de l'utilisateur pour les items blanchis.

3. Mesurer l'erreur commise par l'algorithme lors de la prédiction de la mesure par rapport à la vraie mesure auparavant blanchie ;

4. Calculer la moyenne sur tous les utilisateurs de l'ensemble de validation et tous les items blanchis, en appliquant par exemple la mesure MAE ou *Mean Absolute Errror* (cf. section 3.2.6.3).

La mesure la plus utilisée dans le domaine des systèmes de recommandation est sans conteste la racine carrée de la moyenne des différences au carré ou *Root Mean*

Squared Error (RMSE), sélectionnée notamment pour le fameux Challenge (déjà remporté) *Netflix* [11]. L'erreur absolue moyenne ou MAE représentait avant le challenge *NetFlix*, la métrique d'évaluation le plus répandue pour les systèmes de recommandation. Nous nous concentrerons par la suite sur l'explication de ces deux mesures. Par ailleurs, il semble important de contrôler avant toute chose la dispersion du jeu de données d'entrée avant d'évaluer le système de recommandation [141].

3.2.6.1 Dispersion

La dispersion (*sparsity*) d'un jeu de données représente le ratio de remplissage de la matrice d'usage. Cette métrique est définie comme suit :

$$Sparsity = 1 - \frac{|U|}{|P| \times |C|} \tag{3.9}$$

où $|U|$ correspond à la cardinalité des mesures renseignées, $|P|$ à la cardinalité des items et $|C|$ à la cardinalité des utilisateurs. Les caractéristiques des jeux de données les plus rencontrés dans la littérature sont présentées dans le tableau 3.7 [238].

Jeu de données	Domaine	|C|	|P|	|U|	Dispersion
MovieLens 100K	Films	943	1 682	100 000	93,70 %
MovieLens 1M	Films	6 040	3 900	1 000 000	95,75 %
MovieLens 10M	Films	71 567	10 681	10 000 000	98,69 %
EachMovie	Films	72 916	1 628	2 811 983	97,63 %
NetFlix	Films	480 000	18 000	100 000 000	99,99 %
BX	Livres	278 858	271 379	1 149 780	99,99 %
Jester	Blagues	73 421	101	4 100 000	44,71 %
Entree	Restaurants	50 672	4 160	280 000	99,87 %
Ta-Feng	Vente au détail	32 266	2 012	597 890	99,08 %
B&Q	Vente au détail	607 064	577	3 360 579	99,04 %

TABLEAU 3.7 : Les données pour évaluer les systèmes de recommandation [238]

Le jeu de données *MovieLens* a été collecté à l'aide d'une plateforme réalisée par l'un des pionniers du domaine : le groupe de recherche *GroupLens* [12] de l'université du Minnesota. Le jeu de données *EachMovie* [13] a été développé par le groupe HP/Compaq et malgré son indisponibilité depuis début 2004, il s'avère très utilisé dans les expériences des systèmes de recommandation. Le jeu de données *NetFlix* est apparu avec le challenge *NetFlix* de 2006. Le jeu de données *BX* a été collecté à l'aide d'une plateforme de livres en lignes [314] présentant un volume d'items conséquente. A contrario, le jeu de données *Jester* présente un nombre d'items très restreint [101]. Le jeu de donnés *Entree* contient des historiques de sessions Internet. Enfin, les jeux de données *Ta-Feng* et *B&Q* fournissent un ensemble de transactions d'achats dans le domaine de la vente au détail [134].

11. http://www.netflixprize.com
12. http://www.grouplens.org
13. http://www.grouplens.org/node/76

3.2.6.2 *Root Mean Squared Error*

La mesure RMSE permet de mesurer l'erreur réalisée entre la mesure prédite par le système et la mesure réelle donnée par l'utilisateur [141]. La mesure RMSE est une mesure d'erreur, plus sa valeur est faible, moins l'erreur est importante.

$$RMSE = \sqrt{\frac{\sum_{c,p}(prediction_{c,p} - reel_{c,p})^2}{n}} \tag{3.10}$$

où $reel_{c,p}$ représente la mesure réelle de l'utilisateur c sur l'item p, $prediction_{c,p}$ la mesure prédite par le système et n le nombre total de mesures prédites [227]. L'observation des résultats du challenge *Netflix* permet d'estimer une erreur RMSE dans le cadre des systèmes de recommandation. Le moteur utilisé lors du lancement du challenge en 2006, appelé *Cinematch* avait alors une RMSE de 0,9525. L'objectif de ce challenge était d'améliorer ce score de 10 %. Ce dernier a été atteint en juillet 2009 avec une erreur de 0,8567, c'est-à-dire 10,06 % d'amélioration par rapport au score initial [14]. Ce score a été obtenu avec une combinaison de plus de 100 méthodes de prédiction. On peut difficilement imaginer qu'une telle méthode puisse être déployée dans un contexte industriel. Néanmoins, une RMSE de 0,85 permet de se faire une idée des limites probables des performances des systèmes de recommandation.

3.2.6.3 *Mean Absolute Errror*

La mesure MAE [125, 236] (*Mean Absolute Error*) est également une mesure d'erreur permettant de calculer la moyenne des erreurs absolues entre les mesures prédites et réelles. La mesure MAE se calcule de la manière suivante :

$$MAE = \frac{\sum_{c,p} |prediction_{c,p} - reel_{c,p}|}{n} \tag{3.11}$$

où $reel_{c,p}$ représente la mesure réelle de l'utilisateur c sur l'item p, $prediction_{c,p}$ la mesure prédite par le système et n le nombre total de mesures prédites. Par la suite, la mesure d'erreur NMAE [101] signifiant *Normalized Mean Absolute Error*, permet de normaliser la mesure MAE et de comparer différents algorithmes dans le cas où les échelles de votes ne sont pas les mêmes (les bornes correspondant à u_{min} et u_{max}). Cette métrique se calcule de la façon suivante :

$$NMAE = \frac{MAE}{u_{max} - u_{min}} \tag{3.12}$$

3.2.6.4 *High Mean Absolute Errror*

La mesure HMAE [51], acronyme de *High Mean Absolute Errror* est définie comme étant la mesure MAE obtenue uniquement sur les mesures élevées. Par exemple, sur une échelle de votes de 1 à 5 les votes élevés sont les valeurs 4 et 5, ou dans le cadre

14. Un tableau complet des résultats est disponible sur le site internet de NetFlix à l'adresse suivante : http://www.netflixprize.com//leaderboard.

d'achats, nous considérons un seuil minimum de chiffre d'affaires. Cette métrique évalue la qualité des prédictions véritablement importantes, c'est-à-dire celles qui figureront potentiellement dans les *Top-N* recommandations. En effet, il est relativement peu important que la fonction de prédiction peine à estimer les items de mesures faibles (un film qui ne serait pas apprécié ou un achat qui ne serait pas réalisé), puisque ceux-ci ne seront jamais suggérés par le système et s'avèrent inintéressants pour la satisfaction du fournisseur et la fidélisation de l'utilisateur.

3.2.6.5 Précision et rappel

Dans de nombreux systèmes de recommandation, on ne privilégie pas la prédiction de la note mais plutôt les vraies recommandations des fausses. Lorsqu'on cherche à prédire si un utilisateur est intéressé ou non par un item, quatre possibilités sont offertes (cf. matrice 3.8).

		Réel	
		Consommé	Non consommé
Prédit	Recommandé	Vrai Positif (vp)	Faux Positif (fp)
	Non Recommandé	Faux Négatif (fn)	Vrai Négatif (vn)

TABLEAU 3.8 : Matrice de confusion de la recommandation d'un item à un utilisateur

Dès lors, les mesures de *précision* et de *rappel* peuvent être utilisées pour évaluer les systèmes de recommandation.

DÉFINITION 6

La précision [69] correspond au pourcentage ou au nombre d'items suggérés et s'avérant véritablement pertinents pour l'utilisateur. Par exemple, si l'on considère une liste des Top-N recommandations, la précision correspond à la proportion d'items véritablement consommés, appréciés ou achetés par l'utilisateur courant.

$$Precision = \frac{|vp|}{|vp| + |fp|} \tag{3.13}$$

DÉFINITION 7

Le rappel [194] mesure le nombre de recommandations pertinentes émises au regard du nombre total de recommandations pertinentes. Concrètement, on énumère le nombre d'items dont la mesure associée est non nulle et se retrouvant parmi les items suggérés.

$$Rappel = \frac{|vp|}{|vp| + |fn|} \tag{3.14}$$

DÉFINITION 8

Il existe une combinaison des mesures de précision et de rappel appelée nombre F [69]. F est donnée comme étant égale à :

$$F = \frac{2 \times Precision \times Rappel}{Precision + Rappel} \tag{3.15}$$

Si la précision et le rappel sont exprimés en nombre d'items et que $F = N$, alors la qualité du système est considérée comme parfaite.

Enfin, les courbes ROC [15] permettent d'illustrer la performance d'un système de recommandation binaire. Graphiquement, on représente une courbe illustrant le taux de recommandations correctes (vrais positifs) en fonction du nombre de recommandations incorrectes (faux positifs).

3.2.6.6 Satisfaction des utilisateurs

Swearingen et Sinha ont mis en évidence que les utilisateurs ont confiance dans le système de recommandation lorsque celui-ci recommande des items qu'ils apprécient [266]. Les travaux de Cosley et al. [75] soulignent que la satisfaction de l'utilisateur diminue quand un nombre significatif d'erreurs est produit par le système. Ils favorisent ainsi la mesure de précision sur celle du rappel. Sarwar et al. [246] soulignent que pour tout système de recommandation commercial, le plus important est d'éviter les faux positifs. Ainsi, le niveau de satisfaction des utilisateurs peut être facilement établi [194].

Herlocker et al. [124] ont étudié comment l'explication des recommandations sous forme de texte ou d'image peut aider les utilisateurs à être convaincu du système. Le bandeau de recommandations sur le site d'*Amazon* permet à l'internaute d'élargir sa navigation et le système de recommandation *Genius* d'*Apple* permet quant à lui d'accéder à des titres musicaux ou applications en quelques clics.

Mais la précision des recommandations et leur explication ne sont pas les seuls critères importants d'évaluation d'un système de recommandation. Le temps de calcul est souvent un facteur déterminant. De plus, malgré le fait que la précision des recommandations soit importante, le classement des recommandations et le fait que le système minore les faux positifs est primordial. En effet, cette contrainte entre en conflit avec une autre attente des utilisateurs : il ne suffit pas de recommander aux utilisateurs ce qu'ils attendent. Supposons que l'utilisateur soit un admirateur de Serge Gainsbourg ; si le système lui recommande dix disques de Serge Gainsbourg, il aura l'impression que le système ne lui apporte rien. Par conséquent, le compromis entre précision et diversité des recommandations est primordial pour la satisfaction des utilisateurs [67].

Dans la section suivante, nous approfondissons les techniques utilisées dans le cadre du domaine le plus exploré de la recommandation [80] : le filtrage collaboratif.

15. Receiver Operating Characteristic

3.3 Classification des techniques de filtrage collaboratif

Breese et al. [37] identifient deux classes d'algorithmes de filtrage collaboratif : les algorithmes basés sur la *mémoire* et les algorithmes basés sur un *modèle*. Cette classification s'inspire des travaux réalisés par Castagnos [65]. Les algorithmes basés sur la mémoire fonctionnent sur la base de données entière pour réaliser des prévisions [236, 254]. Les algorithmes à base de modèles utilisent une partie de la base de données pour apprendre un modèle qui est alors utilisé pour prédire [131, 218, 281]. Dans les sous sections suivantes, nous approfondirons particulièrement les algorithmes basés sur des modèles qui s'appuient sur des techniques variées telles que les réseaux bayésiens, les règles d'association ou le *clustering*.

3.3.1 Algorithmes basés sur la mémoire

Les algorithmes basés sur la mémoire [236] ne sont pas restreints à l'ensemble P_t des ressources disponibles à l'instant t, mais utilisent l'ensemble de la matrice d'usage (cf. section 3.2) où chaque ligne correspond à un utilisateur $c_i \in C$ et chaque colonne correspond à un item $p_j \in P$. En effet, les utilisateurs ont pu évaluer des items ne figurant plus dans P_t. Rappelons que la mesure d'intérêt d'un utilisateur c pour un item p est noté $u(p, c)$.
La prédiction pour l'utilisateur c_i sur l'item p_j est donnée par la formule suivante [66] :

$$f_{predict}(p_j, c_i) = \overline{u(c_i)} + \kappa \sum_{c \in C} f_{simil}(c_i, c)[u(p_j, c) - \overline{u(c)}] \tag{3.16}$$

Avec :

- $u(p_j, c)$ l'intérêt mesuré de l'item p_j par l'utilisateur c ;
- $\overline{u(c)}$ la moyenne de l'ensemble des mesures d'intérêts de l'utilisateur c ;
- $f_{simil}(c_i, c)$ le coefficient de pondération liant l'utilisateur c_i à l'utilisateur c ;
- κ un coefficient de normalisation ;
- C l'ensemble des utilisateurs considérés.

L'apprentissage des « communautés » d'utilisateurs se fait implicitement. En effet, il s'agit de tenir d'autant plus compte de l'avis d'un utilisateur qu'il est proche de l'utilisateur courant, ce qui revient à identifier les utilisateurs les plus pertinents (plus proches voisins) sans avoir besoin de les regrouper en communautés.
Les algorithmes basés sur la mémoire présentent plusieurs avantages : la simplicité et l'évolution dynamique en fonction des comportements des utilisateurs. En effet, la mise à jour d'un profil utilisateur se répercute instantanément sur le calcul de la prédiction. Breese et al. [37] s'accordent toutefois à trouver le changement d'échelle problématique : « si ces méthodes fonctionnent bien sur des exemples de tailles réduites, il est difficile de passer à des situations caractérisées par un grand nombre d'items et/ou d'utilisateurs. Ces algorithmes nécessitent trop de temps et de mémoire pour les bases de données volumineuses ».

3.3.2 Algorithmes basés sur un modèle

Les algorithmes basés sur un modèle constituent une alternative à la complexité combinatoire des algorithmes basés sur la mémoire [65]. Ces derniers créent des modèles descriptifs corrélant utilisateurs, items et mesures d'intérêt. Lorsque la corrélation porte sur les utilisateurs, on parle de « communautés d'intérêts ». Lorsqu'il s'agit de rapprocher les items en fonction des usages, i.e des items appréciés ou achetés par plusieurs utilisateurs, on parle de « *clusters d'items* ». Les prédictions sont ensuite inférées depuis ces modèles. Les modèles sont le plus souvent probabilistes ou basés sur des méthodes de *clustering* [238].

3.3.2.1 Approches probabilistes

Les algorithmes basés sur les modèles ne disposent pas forcément de la totalité des informations relatives aux utilisateurs. Par conséquent, le principe est d'évaluer une espérance $E(u(p_j, c_i))$ [37] de la mesure d'un item p_j par un utilisateur c_i ou de rechercher la mesure la plus probable [248].

$$f_{simil}(p_j, c_i) = E(u(p_j, c_i)) = \sum_{j=u_{min}}^{u_{max}} j.P(u(p_j, c_i) = j \mid u(p, c_i), p \in E_{c_i}) \tag{3.17}$$

avec, E_{c_i} l'ensemble des items mesurés par c_i, P la probabilité que l'utilisateur c_i donne la note j à l'item p_j connaissant ses mesures antérieures $u(p, c_i)$. u_{max} correspond à la valeur maximale d'une évaluation, les notes attribuées par les utilisateurs pouvant être réparties de 1 à 5 par exemple. La problématique est alors de construire un modèle à partir duquel l'évaluation de l'espérance peut être résolue.
Dans les parties suivantes, nous présentons les approches les plus courantes basées sur des modèles pouvant revêtir par exemple la forme de réseaux bayésiens, d'arbres de décision ou de règles d'association.

Classifieur naïf de Bayes
L'idée de base s'appuie sur le fait qu'il existe des groupes d'utilisateurs ayant des caractéristiques similaires, i.e ayant un comportement de vote ou d'achat semblable. L'utilisation d'un classifieur naïf de Bayes repose sur l'hypothèse d'indépendance des préférences. Par exemple, on considère que les votes ou les achats de différents items sont indépendants les uns des autres. Le modèle reliant la probabilité jointe d'une classe et des votes à un ensemble malléable de distributions conditionnelles est la formule « naïve » de Bayes [65] :

$$P(C = c, u_{a,1}, \ldots, u_{a,m}) \propto P(C = c) \prod_{p=1}^{m} P(u_{a,p} | C = c) \tag{3.18}$$

Le terme $P(C = c, u_{a,1}, \ldots, u_{a,m})$ correspond à la probabilité d'observer un utilisateur dans une classe c. Les probabilités $P(C = c)$ et $P(u_{a,p}|C = c)$ sont estimées à partir d'un jeu de données d'apprentissage composé de mesures d'utilisateurs. Dans le cas où il n'est pas possible d'observer les variables de classes dans la base

de données utilisateurs, il est possible de recourir à des méthodes apprenant des paramètres pour des modèles présentant des variables cachées. L'algorithme EM (*Expectation Maximization*) permet d'apprendre les paramètres pour une structure de modèle avec un nombre fixé de classes. Le choix du nombre de classes peut être basé sur la mesure de vraisemblance des données [72].
Miyahara et Pazzani [196] proposent de travailler sur une classification des items pour l'utilisateur courant à partir de votes d'autres utilisateurs. Ils définissent deux classes : *aime* ou *n'aime pas*, travaillant ainsi sur une matrice booléenne en fonction d'un seuil fixé empiriquement par l'utilisateur. En l'absence d'information dans la matrice de vote, la valeur est 0. Pour déterminer la classe prédite, la formule 3.18 est adaptée. Une phase d'apprentissage est également requise, déterminant ainsi la classe la plus probable pour l'utilisateur courant.
Le classifieur naïf de Bayes est simple à mettre en œuvre et peu coûteux. Au demeurant, l'hypothèse d'indépendance des mesures d'intérêts s'avère peu réaliste. En effet, prenant l'exemple des ventes croisées, un consommateur aura tendance à acheter un produit complémentaire en fonction d'un autre produit déjà acheté. De la même manière, un cinéphile aura tendance à regarder l'ensemble d'une trilogie.

Arbres de décision et réseaux bayésiens
L'approche proposée par Breese et al. [37] consiste à construire un réseau bayésien dans lequel chaque nœud correspond à un item et l'état du nœud à une mesure d'intérêt. La phase d'apprentissage consiste à rechercher les dépendances entre les items et à construire les tables des probabilités conditionnelles. Par exemple, dans le cas d'achats d'items, les items seront dépendants entre eux s'ils sont achetés par une même population d'utilisateurs. À chaque nœud est associé un item p_k. Ensuite, sont insérées les évidences, c'est-à-dire les mesures des utilisateurs pour les items et/ou les items parents. Dès lors, les tables de probabilités conditionnelles permettent de calculer la probabilité que c_i apprécie l'item p_j.
Il est également possible de représenter chaque table de probabilités conditionnelles sous la forme d'un arbre de décision en fonction des prédécesseurs les plus à même de prévoir l'état du nœud. L'appréciation qui a été faite des prédécesseurs définit le parcours dans l'arbre de décision permettant de prévoir l'intérêt porté par l'utilisateur à l'item courant [72].
Breese et al. [37] démontrent que l'approche à base de réseaux bayésiens fournit des recommandations de très bonne qualité. Par ailleurs, les réseaux bayésiens sont appréciables pour leur capacité à mêler les probabilités issues d'un traitement statistique de retour d'expérience et les probabilités subjectives. Ils permettent également d'expliciter les relations existantes entre les items. Toutefois, le temps de calcul nécessaire lors de la construction du réseau devient totalement inapproprié dans un contexte industriel lorsqu'il y a un très grand nombre d'items.
Heckerman et al. [121] proposent la généralisation de l'approche bayésienne par des réseaux de dépendances, c'est-à-dire une permission des cycles dans le graphe. Les résultats obtenus sont légèrement moins précis qu'avec les réseaux bayésiens mais le besoin en espace mémoire et le temps de calcul sont réduits. Une alternative aux approches probabilistes est abordée dans la section suivante : le *clustering*.

3.3.2.2 *Clustering*

Les méthodes dites de *clustering* permettent de limiter le nombre d'utilisateurs considérés dans le calcul de la prédiction. Le temps de traitement est diminué et les résultats potentiellement plus pertinents puisque les observations portent sur un groupe d'utilisateurs ayant un comportement semblable [116]. Nous verrons dans cette section qu'il est possible de réaliser du *clustering* basé sur les corrélations non seulement entre utilisateurs, mais également entre items [245]. Par exemple, Shani et al. [253] utilisent les listes de films favoris des internautes sur leur blog *MySpace* [16] pour réaliser des recommandations basées sur les corrélations entre items en appliquant sur ces listes des méthodes de *clustering* basées sur les co-occurences de films.

K-Means

La méthode des *K-means* [123] consiste dans un premier temps à choisir aléatoirement k centres dans l'espace de représentation utilisateurs/items. Chaque point de l'espace correspond à un utilisateur dont les coordonnées sont les mesures. Ensuite, chaque utilisateur est positionné dans le *cluster* de centre le plus proche. Une fois les groupes d'utilisateurs formés, la position des centres est recalculée et l'opération réitérée jusqu'à l'obtention d'un état stable. La complexité algorithmique est en $O(knt)$ où k est le nombre de *clusters*, n le nombre d'utilisateurs et t le nombre d'itérations. La prédiction fonctionne pour les algorithme basés sur la mémoire. Au demeurant, la prédiction ne porte plus sur l'ensemble des utilisateurs C mais sur des *clusters* d'utilisateurs appartenant au même groupe que l'utilisateur courant c_i.

D'autres algorithmes proches des *K-means* existent. Par exemple, *Firefly* [254] consiste à sélectionner uniquement les profils utilisateurs dont la valeur de corrélation de *Pearson* par rapport à l'utilisateur courant c_i est supérieure à un seuil fixé empiriquement.

La méthode des *K-means* est utilisée dans de nombreux systèmes de recommandation. Cependant, il est très difficile de connaître le nombre k de centres appropriés. D'autre part, cet algorithme est coûteux en temps de calcul et peut présenter des problèmes de convergence. En effet, les *clusters* générés sont très dépendants de la phase d'initialisation de l'algorithme et il est souvent rare de tomber sur un optimum global minimisant les distances intra-groupes et maximisant les distances inter-groupes. D'une façon générale, il est même fréquent que l'algorithme ne se termine pas. Par ailleurs, les résultats sont non reproductibles, c'est-à-dire que si l'on lance deux fois de suite l'algorithme avec des paramètres identiques, les résultats s'avèrent différents.

Bien que les *K-means* représentent la méthode de *clustering* la plus populaire, il existe des algorithmes concurrents comme *Repeated Clustering*, *Gibbs Sampling* [281] ou *RecTree* [69]. Seule la constitution des *clusters* varie dans ces méthodes, la phase de prédiction des mesures étant souvent similaire.

16. http://www.myspace.com

Repeated Clustering
L'algorithme de *clustering* répété (*Repeated clustering*) [281] consiste à effectuer des affinements successifs des groupes d'utilisateurs. Ainsi, un premier regroupement d'utilisateurs par rapport aux items mesurés permet de définir des classes d'utilisateurs et/ou des classes d'items. Le processus est réitéré sur les classes issues du premier *clustering*. Cette méthode présente un risque de *surgénéralisation*. Au cours des itérations successives, des utilisateurs aux profils et/ou aux comportements différents risquent d'être regroupés

Gibbs Sampling
L'algorithme *Gibbs Sampling* [281] est une méthode d'estimation de paramètres dans un modèle statistique. L'algorithme se déroule en deux phases :

1. Étape d'attribution : l'utilisateur se voit attribuer une classe proportionnellement à une probabilité calculée ;
2. Étape d'estimation : le système estime la classe à laquelle appartient l'utilisateur pour un item courant.

L'algorithme converge mais est extrêmement coûteux en temps de calcul.

RecTree
Le *clustering* hiérarchisé (*RecTree*) [69] cherche à fractionner l'ensemble des utilisateurs en *cliques*. Celles-ci sont hiérarchisées sous forme d'un arbre comme l'illustre la figure 3.12 et la matrice des votes 3.9 correspondante.

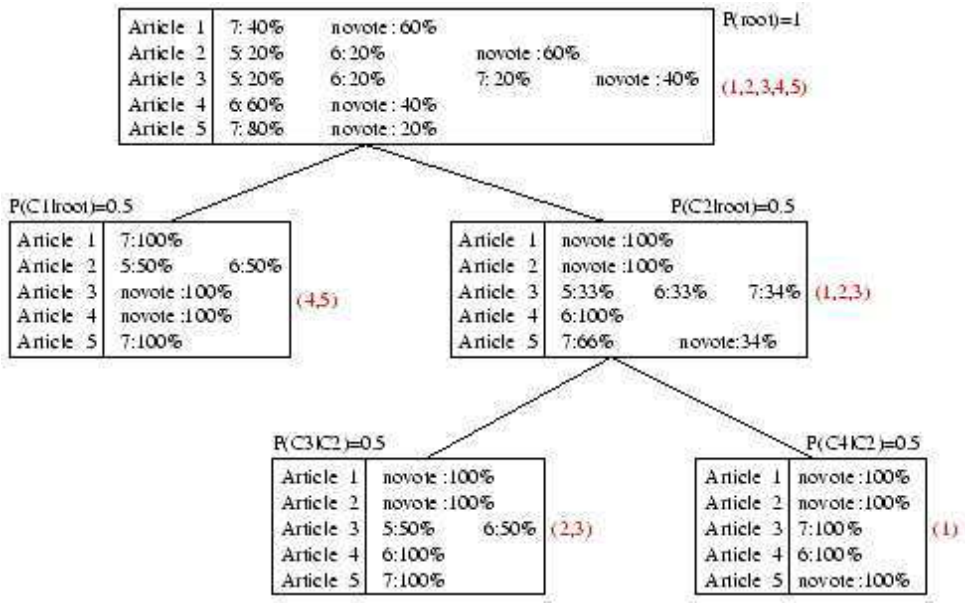


FIGURE 3.12 : Hiérarchie de cliques [65]

	p_1	p_2	p_3	p_4	p_5
c_1			7	6	
c_2			5	6	7
c_3			6	6	7
c_4	7	5			7
c_5	7	6			7

TABLEAU 3.9 : Matrice de votes pour la hiérarchie de cliques [65]

L'algorithme commence par associer à la racine la répartition des mesures de tous les utilisateurs C. Pour construire l'arbre, on cherche à maximiser les similarités entre les utilisateurs d'une même clique et à minimiser celles entre les utilisateurs de deux cliques différentes. Ainsi, plus on descend dans l'arbre et plus les *clusters* sont spécifiques à certains groupes d'utilisateurs similaires. Plus on parcourt l'arbre en profondeur, plus les utilisateurs partagent une mesure semblable sur un item. Sur la figure 3.12, les chiffres à droite de chaque clique indiquent la liste des utilisateurs du groupe. Les probabilités d'atteindre chaque nœud de l'arbre, étant donné le nœud parent sont affichées au-dessus de chaque clique. La probabilité $P(root)$ de la racine de l'arbre vaut 1. Cet algorithme présente les mêmes problèmes de convergence que l'algorithme *K-means*. Il s'avère toutefois intéressant car il permet facilement de subdiviser les communautés d'utilisateurs. Il est potentiellement moins coûteux en temps de calcul que les *K-means* et il possède une moindre complexité combinatoire [65].
Les résultats expérimentés par le *clustering* pour le filtrage collaboratif sont légèrement inférieurs aux résultats fournis par les approches basées sur la mémoire. En effet, les communautés d'utilisateurs créées par le *clustering* se veulent rarement homogènes et certains utilisateurs séparés par les *clusters* apporteraient de l'information pertinente. En revanche, lors de l'utilisation de données volumineuses, le *clustering* permet de franchir le changement d'échelle pour appliquer par la suite une approche basée sur la mémoire.

3.3.2.3 Extraction de règles d'association

L'extraction de règles d'association vise à découvrir des règles de la forme « si un utilisateur c mesure favorablement les deux items p_1 et p_2, alors c mesurera très probablement l'item p_3 ». Dans le contexte des systèmes de recommandation, une transaction peut être considérée comme une évaluation antérieure du client, e.g. un panier d'achats réalisé dans un magasin.
Les règles peuvent être utilisées pour recommander des items en comparant les prémisses des règles de l'utilisateur courant aux prémisses des règles constituant le modèle. Dès lors, la conclusion associée peut être recommandée. Néanmoins, si plusieurs règles sont applicables, les recommandations peuvent être triées en fonction de mesures telles que le support et la confiance considérés comme un score pour la règle. Différentes approches ont été proposées dans la littérature :

- **Approche de la confiance maximale** : seule la confiance maximale de chaque conclusion est conservée. Les conclusions sont ensuite triées par ordre décroissant de confiance [246, 292] ;
- **Approche de la somme des confiances** : pour chaque conclusion de règle, la somme des confiances associées à chaque antécédent est calculée [147] ;
- **Approche de la longueur maximale** : les règles présentant les prémisses les plus longues sont priorisées pour les recommandations [203]. Cette dernière approche pouvant être combinée aux deux premières.

Récemment, les règles d'association sont devenues d'application courante dans les systèmes de recommandation [162]. Leung et al. [161] présentent un modèle basé sur les règles d'association floues en tirant parti des similitudes de produits dans des taxonomies. Liu et al. [176] proposent d'extraire des comportements d'achats de clients à l'aide de règles d'association séquentielles, prenant ainsi en compte le comportement d'achat du client dans le temps. Mobasher et al. [197] présentent un système de personnalisation du Web. Leur système identifie les règles d'association en fonction des co-occurrences rencontrées lors de la navigation des utilisateurs. Smyth et al. [262] présentent deux études de cas de l'utilisation de règles d'association pour les systèmes de recommandation. Lin et al. [168] présentent un nouvel algorithme d'extraction qui ajuste le seuil minimum de support en vue d'obtenir un nombre de recommandations acceptables et utilisables. Cho et al. [73] combinent les arbres de décision et l'extraction de règles d'association. Lawrence et al. [156] présentent un système de recommandation de nouveaux produits aux clients de magasins à l'aide de PDA (*Personal Digital Assistants*). L'auteur caractérise le comportement des clients à l'aide de techniques de *clustering* puis recommande des produits à l'aide de règles d'association. Enfin, Lin et al. [166] et Li et al. [164] se concentrent sur la diminution du nombre de règles d'association pour améliorer le choix et la pertinence des recommandations.
Pour résumer, la plupart des systèmes de recommandation fondés sur des règles d'association se concentrent à recommander un nombre important d'items et se basent principalement sur l'intérêt actuel du client (métaphore du panier de la ménagère) et non sur son historique, s'adaptant seulement à des comportements d'achats éphémères.

3.3.2.4 Approches *Item-Item*

Cette approche s'inspire fortement du principe des ventes croisées. En effet, un utilisateur mesurant un item est généralement intéressé par un autre item. Dès lors, l'algorithme *Item-Item* [245] consiste à construire un modèle identifiant les relations existantes entre les items, à la différence des précédents algorithmes de *clustering* dont le rôle principal était de rapprocher les utilisateurs semblables.
Une manière simpliste d'élaborer un tel type de modèle consiste à compter le nombre de fois qu'une paire d'items est parcourue par un même utilisateur. Par exemple, la matrice 3.10 souligne que les personnes ayant regardé le film *Avatar* ont également

regardé *Inception*. Il peut donc s'avérer pertinent de recommander *Inception* à un utilisateur ayant visualisé *Avatar*. Ce type de modèle fait également ressortir qu'une grande majorité des utilisateurs qui ont regardé *Matrix* 1 ont également regardé *Matrix* 2 *et* 3. Certains items sont corrélés de ce fait, sans qu'il soit nécessaire d'utiliser des méthodes statistiques.

	Avatar	**Inception**	**Matrix 1**	**Matrix 2**	**Matrix 3**
Avatar	-	40	7	6	9
Inception	40	-	5	6	7
Matrix 1	7	5	-	30	29
Matrix 2	6	6	30	-	29
Matrix 3	7	6	29	29	-

TABLEAU 3.10 : Modèle *Item-Item* pour des films co-visionnés

Une fois le modèle réalisé (cf. matrice 3.10), il est possible de recommander à l'utilisateur courant des items corrélés à ceux qu'il a mesurés. Dans notre exemple, il suffit de sélectionner les lignes d'items regardées par l'utilisateur, puis en additionnant colonne par colonne les valeurs des lignes, on obtient un vecteur de corrélation. Dès lors, ce vecteur peut être trié pour obtenir la liste des *Top-N* recommandations pour l'utilisateur courant, correspondantes aux valeurs les plus élevées. Cependant, cette approche présente deux inconvénients majeurs [65] :

1. Les recommandations correspondent aux items les plus « populaires » (cf. phénomène de la longue traîne dans la section 3.2.5.3) ;

2. Il est difficile d'estimer l'intérêt que peut représenter une recommandation par rapport à une autre (absence de prédiction) pour un utilisateur.

Par conséquent, Sarwar et al. [245] ont proposé de construire le modèle de corrélation non plus en fonction du comptage des co-occurences d'items, mais sur la base des mesures fournies par les utilisateurs sur les items. Il s'agit alors de transformer la matrice d'usage U des mesures utilisateurs/items en une matrice U' de similarités entre les items. La matrice générée est symétrique et la distance utilisée est souvent le vecteur *cosine* ajusté [17]. Sarwar et al. avancent l'hypothèse qu'il est plus aisé de rapprocher les items que les utilisateurs car il existe potentiellement moins d'items que d'utilisateurs et parce que la liste des items P est souvent plus « stable ». Une approche basée sur les items s'avère plus pertinente lorsque le nombre d'utilisateurs $|C|$ est très supérieur au nombre d'items $|P|$, c'est-à-dire $|C| \gg |P|$.
Par ailleurs, l'algorithme *Item-Item* produit des prédictions de très bonne qualité bien que les raisons avancées par Sarwar et al. soient discutables. En effet, selon le contexte applicatif, il n'est pas rare que le nombre d'items dépasse fortement le nombre d'utilisateurs. Par exemple, un magasin de ventes au détail présentera souvent des milliers d'items alors que le nombre de clients est souvent de quelques centaines, parfois de quelques milliers. De même, la stabilité de l'ensemble des items

17. Une adaptation du coefficient de *Pearson*

peut être remise en cause. Sur le Web, les items vendus sur un site marchand sont souvent les mêmes alors que dans la vente au détail de matériaux de construction, les items peuvent être extrêmement volatiles. Le choix d'une méthode basée sur les utilisateurs ou les items dépend fortement du domaine applicatif [66].

3.3.3 Algorithmes basés sur la mémoire et sur un modèle

Pennock et al. [218] s'accordent à penser que les algorithmes basés sur un modèle ont une moindre complexité combinatoire que ceux basés sur la mémoire. Toutefois, ils présentent globalement une évolutivité trop lente : la mise à jour du modèle est en effet coûteuse en temps de calcul. L'hybridation s'attache à combiner différentes approches des filtrages collaboratifs et thématiques. Burke [49] présente sept techniques principales d'hybridation :

1. *Weighted* : interpolation des scores des différentes recommandations ;
2. *Switching* : choix de la technique de recommandation la plus appropriée ;
3. *Mixed* : concaténation des recommandations issues de techniques différentes ;
4. *Feature Combination* : utilisation de la combinaison d'attributs provenant de techniques différentes ;
5. *Feature Augmentation* : utilisation d'une première technique pour la calcul d'attributs qui sont ensuite utilisés dans une seconde technique ;
6. *Cascade* : instauration d'une hiérarchie au sein des modèles, les moins haut placés servant à renforcer les scores de ressources obtenus avec les modèles les plus haut placés ;
7. *Meta-level* : une première technique construit un modèle utilisé par une seconde technique.

Des algorithmes hybrides ont vu le jour, cherchant à combiner les avantages des approches basées sur la mémoire (réactivité et qualité des prédictions) de celles basées sur un modèle (moindre complexité) [65, 141].

3.3.3.1 *Horting*

Le filtrage collaboratif *Horting* [3] est une approche basée sur un graphe de relation de similarité (arcs) entre les utilisateurs (nœuds). La notion d'influence se décline sous forme de deux contraintes :

1. La contrainte de *Horting* impose de ne considérer que les utilisateurs ayant un grand nombre de mesures communes ;
2. La contrainte de prédictabilité ajoute à la notion de *Horting* une information sur le degré de ressemblance entre deux utilisateurs en se basant sur la distance de *Manhattan*.

Le parcours du graphe permet de filtrer les utilisateurs proches et ceux ayant un nombre de mesures important. La notion de prédictabilité est plus contraignante que le concept de proximité car le système a besoin d'un échantillon suffisamment important de ressources communément mesurées.

3.3.3.2 *Eigentaste*

L'approche basée sur le *clustering* et la réduction de dimension, appelée *Eigentaste* [101] se décompose en trois étapes :

1. Chaque utilisateur commence par évaluer un ensemble d'items « jauges » décrit par des métadonnées ;

2. La deuxième phase débute par une Analyse en Composantes Principales (ACP) afin d'obtenir une réduction optimisée de l'espace de représentations utilisateurs/items. Puis ce même espace réduit est partitionné selon le principe de *clustering* rectangulaire récursif. Pour chaque groupe créé, le système calcule la moyenne des mesures autres que celles de l'ensemble de la jauge. Une table de recommandations est ensuite initialisée avec la liste des moyennes calculées, triées par ordre croissant ;

3. Enfin, la dernière étape consiste à rechercher le *cluster* le plus approprié et à récupérer la prédiction dans la table des recommandations afin d'obtenir l'ensemble des items associés au *cluster*.

Cette méthode présente deux inconvénients majeurs. La première phase est contraignante pour l'utilisateur qui doit jauger un ensemble d'items et la deuxième phase peut engendrer des problèmes de rafraîchissement des informations.

3.3.3.3 Diagnostic de personnalité

Pennock et al. [218] ont introduit une méthode de modélisation de la personnalité permettant de considérer les mesures des utilisateurs avec un bruit Gaussien. Ainsi, il est possible de prendre en compte des paramètres extérieurs tels que le caractère de l'utilisateur par exemple. Les mesures sont considérées comme des symptômes et le type de personnalité comme une maladie. De cette manière, déterminer cette personnalité revient à identifier la cause la plus probable des mesures. Cet algorithme, bien que basé sur un modèle probabiliste, se gère comme une approche basée sur la mémoire car toutes les informations sont nécessaires pour les calculs de probabilité. Le principal bénéfice de cette approche réside dans le fait que les prédictions du modèle sont exprimées explicitement, rendant ainsi possible leur modification et leur validation.
L'efficacité des modèles hybrides a émergé lors du concours *NetFlix*. En effet, le meilleur résultat a été obtenu par une approche hybride mélangeant pas moins de 107 modèles [25], ces derniers représentant principalement des variantes des cinq modèles de base.

3.3.4 Synthèse de la classification

Nous présentons dans le tableau 3.11 une synthèse des approches de filtrage collaboratif [65, 141, 227, 238].

	Techniques de filtrage collaboratif		
	Approches basées sur la mémoire	**Approches basées sur un modèle**	**Approches basées mémoire et modèle**
Techniques	▶ Plus proches voisins	▶ Réseaux Bayésiens ▶ Arbres de décisions ▶ *Clustering* ▶ Règles d'association ▶ Approche *Item-Item*	▶ *Horting* ▶ *Eigentaste* ▶ Diagnostic de personnalité
Avantages	▶ Réactif ▶ Simpliste ▶ Performance	▶ Raisonnement Prédictif ▶ Non dynamique	▶ Combinaison des deux approches
Inconvénients	▶ Complexité combinatoire	▶ Moindre complexité	▶ Multiplication des modèles

TABLEAU 3.11 : Synthèse des techniques de filtrage collaboratif

- Les algorithmes basés sur la mémoire offrent l'avantage d'être réactifs, en intégrant dynamiquement des nouveaux utilisateurs ou items. Au demeurant si ces méthodes fonctionnent bien sur des exemples de tailles réduites, il est souvent difficile de passer à des situations proposant un grand nombre d'items ou d'utilisateurs, du fait notamment à la complexité combinatoire des algorithmes utilisés.

- Les algorithmes basés sur un modèle offrent une valeur ajoutée au-delà de la seule fonction de prédiction. En effet, ils mettent en lumière certaines corrélations dans les données, proposant ainsi un raisonnement intuitif pour les recommandations ou rendant simplement les hypothèses plus explicites. Une autre manière d'aborder le problème du filtrage collaboratif consiste à classifier les utilisateurs et les items en groupes. Pour chaque groupe d'utilisateurs, il s'agit d'estimer la probabilité qu'un item soit choisi. Ces approches souffrent bien souvent de problèmes de convergence liés à l'initialisation des *clusters* et fournissent dans certains cas des recommandations de mauvaise qualité. Les algorithmes basés sur un modèle minimisent le problème de la complexité combinatoire. Cependant, ces méthodes ne sont pas assez dynamiques et elles réagissent mal à l'insertion de nouveaux contenus dans la base de données.

- Les algorithmes basés sur la mémoire et sur un modèle offrent une alternative combinant les avantages des deux approches : réactivité et qualité des prédictions de l'approche mémoire et moindre complexité de l'approche modèle.

À la lecture de cet état de l'art, la problématique principale du filtrage collaboratif reste le changement d'échelle des systèmes. Une solution consiste à scinder de manière explicite les calculs dits « on-line » de ceux dits « off-line » [66].

3.4 Domaines d'applications

Il existe de nombreux systèmes collaboratifs développés autant dans le monde industriel que dans le monde académique. Le système *Grundy* [239] était le premier système de recommandation proposant d'utiliser les stéréotypes en tant que mécanisme pour la construction de modèles. Plus tard, le système *Tapestry* [100] s'est appuyé sur chaque client pour identifier les clients partageant les mêmes idées. *GroupLens* [153], *Video Recommender* [128], et *Ringo* [254] utilisent également des algorithmes de filtrage collaboratif. Nageswara et Talwar [201] proposent une classification des systèmes de recommandation en six catégories suivant la fonctionnalité à laquelle ils répondent :

1. *Content-based filtering systems* : utilisant les données sur les items et le profil de l'utilisateur courant ;
2. *Collaborative filtering systems* : utilisant des données sur un ensemble de comportements utilisateurs interagissant avec un item ;
3. *Demographic filtering systems* : utilisant des données démographiques telles que l'âge, le sexe, le niveau social, etc. permettant de segmenter des populations les rapprochant de certains items ;
4. *Knowledge-based recommender systems* : utilisant de la connaissance fonctionnelle pour générer des recommandations ;
5. *Utility-based recommender systems* : utilisant une fonction d'utilité sur les items pour aider à la recommandation ;
6. *Hybrid recommender systems* : utilisant plusieurs approches pour minimiser les inconvénients de certaines méthodes.

Montaner et al. [198] produisent une taxonomie et classifient les systèmes de recommandation existants en plusieurs domaines :

- Les divertissements (*entertainment*) : films, musiques, etc. ;
- Les contenus (*content*) : actualités personnalisées, pages Web, applications de *e-learning*, antispams, etc. ;
- Le commerce électronique : livres, appareils photos, ordinateurs, etc. ;
- Les services (*services*) : voyages, expertises, locations, etc.

Ricci et al. [238] présentent une classification des domaines de recommandations existants en fonction de plusieurs critères d'évaluation subjectifs dont le risque d'impact sur le client suite à une mauvaise recommandation. Il constate par exemple, que les sites d'assurance vie, de tourisme et de recherche d'emplois ont davantage de risque que les sites de commerce électronique, d'actualités, de films ou de musiques.

Les systèmes de recommandation sont vitaux pour les sites de commerce en ligne, dont les exemples les plus frappant sont *Amazon*, *NetFlix*, *Pandora* et *Strands*. Les systèmes de recommandation touchent principalement aujourd'hui quatre domaines commerciaux en ligne : les **films** [18], la **musique** [19], les **livres** [20] et la **publicité** [21].

La recherche dans le domaine des systèmes de recommandation en *m-commerce* [22] s'est d'ailleurs accélérée ces dernières années. Dans ce cadre, les applications sont nombreuses et variées, nous pouvons mentionner par exemple le tourisme [287] ou la recommandation dans le domaine de la restauration [132]. Le *m-commerce* ouvre des perspectives de recherche autour de la mobilité, de la capacité de calcul limitée, des capacités de transmission, de la taille de l'écran, etc.

Le tableau 3.12 présente une liste non exhaustive d'exemples de systèmes de recommandation commerciaux et académiques, leur domaine d'application et la technique de filtrage utilisée.

		Systèmes collaboratifs		
Système	**Domaine**	**Thématique**	**Collaboratif**	**Hybride**
Adaptive Place [270]	Restaurants		√	
Amazon [169]	Livres, films, etc.			√
Eigenstate [101]	Académique		√	
Fab [16]	Livres			√
InfoFinder [154]	Actualités	√		
Last.fm [122]	Musique			√
LIBRA [30]	Livres			√
Google News [80]	Actualités	√		
GroupLens [153]	Actualités		√	
MovieLens [124]	Films		√	
MYCIN [44]	Prescriptions		√	
Netflix [25]	Films			√
Org. Structure [228]	Appareils photos		√	
Pandora [43]	Musique		√	
RecTree [69]	Images			√
Ringo [47]	Musique		√	
Tapestry [100]	Images		√	
SASY [78]	Vacances		√	
Top Case [186]	Vacances		√	
TrustWalker [139]	Académique	√		

TABLEAU 3.12 : Classification des systèmes collaboratifs commerciaux et académiques

18. http://www.netflix.com
19. http://www.last.fm
20. http://www.amazon.com
21. http://www.facebook.com
22. Les applications en *m-commerce* ne couvrent pas seulement les applications du e-commerce, mais également les nouvelles applications qui peuvent être exécutées à tout moment, de n'importe quel endroit via les mobiles ou les tablettes [135].

3.5 Positionnement

Pour mettre en œuvre les connaissances dites actionnables, nous avons fait le choix des systèmes de recommandation. Ces derniers permettent de réaliser des recommandations personnalisées aux clients via des commerciaux. La plupart des systèmes de recommandation existants se concentrent à prédire la note d'un utilisateur pour un item. En revanche, peu de travaux introduisent une sémantique économique à la mesure d'utilité d'un utilisateur pour un item. En effet, peu de systèmes de recommandation cherchent à prédire les clients qui peuvent dépenser ou dépenser plus pour un produit. L'approche du filtrage collaboratif basée sur un modèle permet de s'appuyer des techniques de fouille de données pour prédire les valeurs de la matrice d'usage, par exemple l'extraction des règles d'association. Les règles d'association représentent un modèle explicite pouvant aider les commerciaux à développer la valeur de leurs clients. Effectuer des recommandations dites intrusives sur le terrain nécessite d'éviter les fausses recommandations. En effet, les mauvaises recommandations peuvent avoir un impact important sur la fidélisation des clients et sur l'adhésion du commercial. Améliorer la mesure de précision vise à réduire le nombre de faux positifs tout en maintenant le nombre de vrais positifs.

3.6 Conclusion

Dans ce chapitre, nous avons décrit les différents systèmes collaboratifs existants : à savoir les filtrages thématique et collaboratif. Le tableau 3.13 illustre les avantages et les inconvénients de ces deux approches. Étant donné l'avantage industriel du filtrage collaboratif, nous avons développé l'explication des techniques les plus couramment utilisées dans la section 3.3.

	Filtrage thématique	**Filtrage collaboratif**
Données	▶ Descriptives, profils	▶ Matrice d'usage
Avantages	▶ Petit volume ▶ Items court cycle de vie	▶ Performance ▶ Items durables
Inconvénients	▶ Moins performant ▶ Problème de l'entonnoir	▶ Grand volume ▶ Problème du démarrage à froid ▶ Problème de la longue traîne

TABLEAU 3.13 : Comparaison des filtrages thématique et collaboratif

Dans la suite de la thèse, nous nous intéresserons principalement au filtrage collaboratif et aux recommandations personnalisées qui répondent à une stratégie commerciale basée sur les forces de vente. En effet, les systèmes collaboratifs s'adressent davantage au monde industriel qui manipule couramment de grands volumes et qui souhaite de plus en plus analyser le comportement de leurs clients. Le filtrage collaboratif semble être la méthode de recommandation personnalisée qui garantit les meilleurs résultats. Dans le prochain chapitre de cette thèse, nous présentons notre méthodologie pour les systèmes de recommandation, fondée sur l'analyse des chiffres d'affaires des clients à différents niveaux d'une taxonomie produits.

4

CAPRE : une méthodologie de recommandations actionnables et profitables

The recommendations provided are aimed at supporting their users in decision-making business processes [...]

Francesco Ricci, 2011

SOMMAIRE

Les chapitres bibliographiques 2 et 3 mettent en évidence le verrou scientifique de l'actionnabilité en ECD, et le verrou applicatif de l'adaptation des systèmes de recommandation pour les commerciaux. Néanmoins, l'actionnabilité est cruciale pour mettre en œuvre des actions issues des modèles et nécessite d'être complétée par le profit. Les systèmes de recommandation adaptés aux commerciaux doivent veiller à minimiser les fausses recommandations. L'extraction des règles d'association est une technique explicite permettant de développer la valeur client, dès lors qu'elle s'appuie sur les systèmes de recommandation. Dans ce chapitre, nous proposons une nouvelle méthodologie *CAPRE* (*Customer Actionability and Profitability REcommendation*) fondée sur l'extraction de règles d'association pour les systèmes de recommandation. Notre méthodologie, instanciée dans le chapitre 6, est décomposée en sept étapes successives et complémentaires présentées sur la figure 4.1.

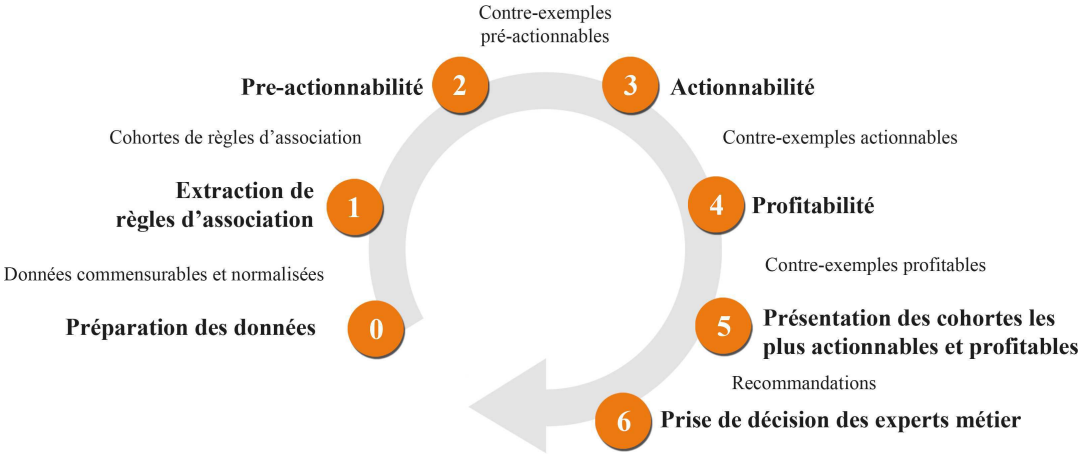


FIGURE 4.1 : Les cinq étapes clefs de la méthodologie *CAPRE*

Étape 0 préparation des données (cf. section 4.2) ;

Étape 1 extraction de comportements d'achats sous forme de règles (cf. section 4.3) ;

Étape 2 mesure de la pré-actionnabilité des contre-exemples sur les variables d'achats (cf. section 4.4.1) ;

Étape 3 mesure de l'actionnabilité des contre-exemples sur les variables descriptives (cf. section 4.4.2) ;

Étape 4 mesure de l'intérêt économique des contre-exemples actionnables (cf. section 4.5) ;

Étape 5 présentation des cohortes les plus actionnables et profitables (cf. section 4.6) ;

Étape 6 prise de décision des experts métier en fonction des contraintes du domaine (cf. section 4.7).

Une phase d'expérimentation et de validation est présentée dans le chapitre 5 et une application sur les données réelles de VM Matériaux est proposée dans le chapitre 6.

4.1 Terminologie et notations

Nous considérons l'ensemble C de tous les clients $\{c_1, c_2, \dots c_n\}$ et P l'ensemble de tous les produits $\{p_1, p_2, \dots p_m\}$ qui peuvent être recommandés. Soit U un ensemble ordonné mesurant l'utilité portée par le client $c \in C$ à l'item $p \in P$. Les principales notations utilisées dans le chapitre 4 sont résumées dans le tableau 4.1.

Notation	Signification
C	un ensemble de n clients notés $\{c_1, c_2, \dots c_n\}$
P	un ensemble de m produits notés $\{p_1, p_2, \dots p_m\}$
N	variables descriptives numériques centrées réduites des clients
I	indicatrices réduites des variables descriptives catégoriques des clients
$u(p,c)$	une fonction d'utilité du produit p pour le client c
$u^{\%}(p,c)$	une version normalisée de la fonction d'utilité u
$u^{*}(p,c)$	une version discrétisée de la fonction $u^{\%}$
U	matrice des valeurs prises par u
R	l'ensemble des règles d'association
$Ct(p)$	une cohorte de règles pour un produit p
$Ct^{+}(p)$	l'ensemble des exemples de la cohorte $Ct(p)$
$Ct^{-}(p)$	l'ensemble des contre-exemples de la cohorte $Ct(p)$

TABLEAU 4.1 : Résumé des notations

Tout au long de la méthodologie, nous illustrerons les différents concepts abordés au travers d'un exemple composé d'achats en Euros de dix clients $\{c_1, c_2, \dots c_{10}\}$ pour cinq produits $\{p_1, p_2, \dots p_5\}$ sur une période prédéterminée (cf. tableau 4.2). Par exemple, le client c_2 a réalisé 2 900 € de chiffre d'affaires (CA) sur le produit p_4. Les cases vides de la matrice 4.2 sont considérées comme des « non achats » des produits (c'est-à-dire des valeurs nulles ou trop faibles).

	p_1	p_2	p_3	p_4	p_5
c_1	2 500 €			3 600 €	17 000 €
c_2	500 €			2 900 €	8 500 €
c_3	1 000 €			2 000 €	300 €
c_4	1 500 €	100 €	200 €	1 500 €	
c_5	950 €	920 €	930 €	990 €	2 400 €
c_6		7 000 €	14 000 €		500 €
c_7		2 000 €	3 000 €		
c_8		2 500 €		1 000 €	100 €
c_9	4 000 €	3 000 €	2 600 €	6 000 €	600 €
c_{10}	900 €				22 800 €

TABLEAU 4.2 : Matrice d'usage des $u(p,c)$ de 10 clients pour 5 produits

Dès lors, nous calculons la dispersion (*sparsity*) du jeu de données (cf. section 3.2.6.1) :

$$Sparsity = 1 - \frac{|U|}{|P| \times |C|} = 1 - \frac{33}{5 \times 10} = \mathbf{0{,}34} \tag{4.1}$$

Le taux de remplissage de la matrice est important. En effet, sur un jeu de données réel, il est rare de voir une dispersion inférieure à 0,95. Ceci s'explique par le fait que les clients achètent généralement très peu de produits différents par rapport au nombre total de produits proposés dans un magasin.

4.2 Préparation des données (étape N° 0, figure 4.1)

Nous définissons une fonction d'utilité u, donnée par la matrice d'usage clients $\times$ produits (cf. tableau 4.2), mesurant l'intérêt d'un produit p pour un client c :

$$\begin{cases} u : P \times C & \rightarrow & \mathbf{R} \\ p, c & \rightarrow & u(p,c) \end{cases} \tag{4.2}$$

où $\mathbf{R}$ est un sous ensemble de $\mathbb{R}$, par exemple les entiers positifs.
Dans notre méthodologie, la fonction u supporte un intérêt économique ou financier : le chiffre d'affaires ou la marge nette d'un produit p par un client c. Sémantiquement, plus la fonction $u(p,c)$ est élevée et plus l'achat du produit p par le client c est profitable pour l'entreprise. Cette démarche est directement applicable au contexte usuel des jeux de données de votes (cf. expérimentations du chapitre 5 sur les données *MovieLens*). Dans la suite du chapitre, nous considérons que la fonction d'utilité u désigne un CA (correspondant au cadre applicatif présenté dans le chapitre 6).
Avant de pouvoir extraire des règles de comportements sur ces données, deux transformations doivent être appliquées sur les valeurs de la fonction d'utilité (cf. figure 4.2).

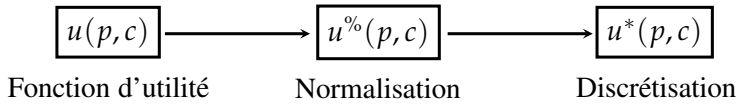


FIGURE 4.2 : Normalisation et discrétisation de la fonction d'utilité

- La première étape consiste à rendre commensurable les chiffres d'affaires d'un client c à un autre, indépendamment de leurs capacités de dépenses. Pour chaque client c, les CA en valeur effective pour chaque produit sont transformés en pourcentage du CA total du client, représentant ainsi la « part » d'utilité d'un produit p pour un client c.

$$\begin{cases} u^{\%} : P \times C & \rightarrow & [0; 100\ \%] \\ p, c & \rightarrow & u^{\%}(p,c) \end{cases} \tag{4.3}$$

avec,

$$u^{\%}(p,c) = \frac{u(p,c)}{\sum_{p \in P} u(p,c)} \tag{4.4}$$

- La deuxième étape consiste à normaliser les proportions $u^{\%}(p,c)$ en prenant en compte leur distribution pour chaque produit. Cette normalisation se traduit par une discrétisation des proportions en plusieurs tranches $\mu_1, \mu_2 \ldots \mu_q$. Pour simplifier les notations, nous considérons que la discrétisation est la même pour tous les produits, ce qui ne nuit pas à la généricité de l'approche. Dans les applications décrites dans les parties 5.2 et 6.2, nous choisissons une discrétisation spécifique pour chaque produit.

$$\begin{cases} u^* : P \times C & \rightarrow & \Omega = \{\mu_1, \mu_2, \ldots \mu_q\} \\ p, c & \rightarrow & u^*(p, c) \end{cases} \tag{4.5}$$

où $\mu_1 < \mu_2 \ldots < \mu_q$. La normalisation et la discrétisation dépendent fortement du contexte applicatif.

EXEMPLE 3 Reconsidérons le tableau 4.2 présentant la matrice d'utilité de dix clients pour cinq produits. Normalisons tout d'abord les valeurs $u(p, c)$ pour obtenir les pourcentages $u^{\%}(p, c)$ (cf. tableau 4.3). Ensuite, appliquons la fonction de discrétisation $u^*(p, c)$ (cf. tableau 4.4). Nous illustrons avec une discrétisation sur trois intervalles distincts et de même fréquence dépendant du produit p. Pour chaque produit p, les intervalles $a(p)$, $b(p)$ et $c(p)$ identifient trois niveaux de part de chiffre d'affaires : faible, moyen et fort. $u^*(p, c) = a$ (respectivement b et c), signifiant que le chiffre d'affaires du client c pour le produit p est faible (respectivement moyen et fort).

	p_1	p_2	p_3	p_4	p_5
c_1	10,82 %			15,58 %	73,60 %
c_2	4,21 %			24,37 %	71,42 %
c_3	30,30 %			60,60 %	9,10 %
c_4	45,45 %	3.03 %	6.07 %	45,45 %	
c_5	15,35 %	14,86 %	15,02 %	15,99 %	38,78 %
c_6		32,56 %	65,12 %		2,32 %
c_7		40,00 %	60,00 %		
c_8		69,44 %		27,78 %	2,78 %
c_9	24,69 %	18,52 %	16,05 %	37,04 %	3,70 %
c_{10}	3,80 %				96,20 %

TABLEAU 4.3 : Matrice d'usage des $u^{\%}(p, c)$

	p_1	p_2	p_3	p_4	p_5
c_1	a			a	c
c_2	a			a	b
c_3	b			c	b
c_4	c	a	a	b	
c_5	b	a	a	a	b
c_6		b	c		a
c_7		b	b		
c_8		c		b	a
c_9	b	a	b	b	a
c_{10}	a				c

TABLEAU 4.4 : Matrice d'usage des $u^*(p, c)$

Pour chaque produit, les intervalles de part de chiffre d'affaires correspondants sont exprimés dans le tableau 4.5. Par exemple, les forts chiffres d'affaires pour le produit p_5, i.e $c(p_5)$ correspondent à l'intervalle [72,51 ; 100,00 %].

	p_1	p_2	p_3	p_4	p_5
a	[3,80 ; 13,08]	[3,03 ; 25,54]	[6,07 ; 15,53]	[15,58 ; 26,07]	[2,32 ; 6,40]
b	]13,08 ; 37,87]	]25,54 ; 54,72]	]15,53 ; 62,56]	]26,07 ; 53,02]	]6,40 ; 72,51]
c	]37,87 ; 100]	]54,72 ; 100]	]62,56 ; 100]	]53,02 ; 100]	]72,51 ; 100]

TABLEAU 4.5 : Les trois intervalles $a(p)$, $b(p)$ et $c(p)$ pour chaque produit p

4.3 Extraction de comportements d'achats sous forme de cohortes (étape N° 1, figure 4.1)

Nous appliquons un algorithme d'extraction de règles d'association [68, 114] pour extraire des règles de comportements d'achats de la forme :

$$u^*(p_X, \centerdot) = \mu_X,\ u^*(p_Y, \centerdot) = \mu_Y,\ \ldots \rightarrow\ u^*(p_Z, \centerdot) = \mu_Z \tag{4.6}$$

où $\forall_i\ p_i \in P$ et $\forall_j\ \mu_j \in \Omega$. Les exemples de la règle sont les clients qui satisfont la prémisse et la conclusion (correspondant aux points "$\centerdot$" dans la formule 4.6). A contrario, les clients contre-exemples satisfont la prémisse mais non la conclusion à hauteur du chiffre d'affaires souhaité. Plus précisément, dans la méthodologie *CAPRE*, les clients contre-exemples « éligibles » sont évidemment les clients qui achètent p_Z à un niveau d'achat strictement inférieur à μ_Z. Une telle règle signifie que les clients qui satisfont la prémisse « ont tendance » à satisfaire la conclusion. Notre objectif étant de de recommander un item, notre méthodologie utilise les règles avec un seul élément en conclusion. Par souci de simplification, nous noterons $u^*(p_X, c) = \mu_X$ par $p_X = \mu_X$ (où $p_X \in P$ et $\mu_X \in \Omega$). Par conséquent, la règle présentée dans la formule 4.6 est simplifiée de la manière suivante :

$$p_X = \mu_X,\ p_Y = \mu_Y\ \ldots \rightarrow\ p_Z = \mu_Z \tag{4.7}$$

Dans la terminologie usuelle des règles d'association, $p_X = \mu_X$ est appelé un *item*. L'ensemble des règles d'association, noté R, est extrait en fonction de deux seuils *minSup* et *minConf* déterminés par les experts métier : le support et la confiance respectivement.

DÉFINITION 9

Soit un item $p_X = \mu_X$, une cohorte Ct est l'ensemble des règles d'association concluant sur ce même item :

$$Ct(p_X = \mu_X) = \{r \in R \mid conclusion(r) = (p_X = \mu_X)\} \tag{4.8}$$

où $conclusion(r)$ est la fonction qui donne la conclusion d'une règle r. L'ensemble des règles R est partitionné en autant de cohortes qu'il existe de conclusion de règles.

DÉFINITION 10

Étant donné une cohorte $Ct(p_X = \mu_X)$, les ensembles d'exemples et de contre-exemples, notés $Ct^+(p_X = \mu_X)$ et $Ct^-(p_X = \mu_X)$ sont les unions des ensembles d'exemples et de contre-exemples respectifs des règles d'association appartenant à la cohorte.

Le fait de détecter des groupes de clients contre-exemples de règles peut théoriquement permettre à l'entreprise de recommander certains items et ainsi générer des bénéfices pouvant être estimés *a priori*. L'intérêt de former des cohortes de règles d'association est de fusionner toutes les règles qui auraient tendance à recommander le même item. En exploitant la redondance existante, nous sommes capables d'extraire des comportements d'achats plus robustes, c'est-à-dire défendus par davantage d'exemples.

DÉFINITION 11
Étant donné une cohorte Ct, une Ct-Recommandation consiste à proposer la conclusion de la cohorte Ct à un contre-exemple appartenant à Ct^{-}.

EXEMPLE 4 À l'aide d'un support *minSup* égal à 20 % (2 clients) et d'une confiance *minConf* égale à 50 %, nous pouvons extraire un ensemble de seize règles d'association (cf. tableau 4.6) à partir de la matrice d'usage discrétisée (cf. tableau 4.4).

N° Règle	Support (%)	Confiance (%)	Règle
1	0,2	0,666667	p_1=a $\rightarrow$ p_4=a
2	0,2	0,666667	p_4=a $\rightarrow$ p_1=a
3	0,2	0,666667	p_4=a $\rightarrow$ p_5=b
4	0,2	0,666667	p_5=b $\rightarrow$ p_4=a
5	0,2	1	p_5=c $\rightarrow$ p_1=a
6	0,2	0,666667	p_1=a $\rightarrow$ p_5=c
7	0,2	0,666667	p_5=b $\rightarrow$ p_1=b
8	0,2	0,666667	p_1=b $\rightarrow$ p_5=b
9	0,2	0,666667	p_1=b $\rightarrow$ p_2=a
10	0,2	0,666667	p_2=a $\rightarrow$ p_1=b
11	0,2	1	p_3=a $\rightarrow$ p_2=a
12	0,2	0,666667	p_2=a $\rightarrow$ p_3=a
13	0,2	0,666667	p_2=a $\rightarrow$ p_4=b
14	0,2	0,666667	p_4=b $\rightarrow$ p_2=a
15	0,2	0,666667	p_4=b $\rightarrow$ p_5=a
16	0,2	0,666667	p_5=a $\rightarrow$ p_4=b

TABLEAU 4.6 : Extraction de règles d'association, $minSup = 20$ % et $minConf = 50$ %

Dès lors, neuf cohortes peuvent être générées, regroupant ainsi les règles concluant sur le même item (cf. tableau 4.7).

N° Cohorte	Conclusion	N° des Règles	Contre-exemples	Exemples
1	p_1=a	2, 5	c_5	c_1, c_2, c_{10}
2	p_4=a	1, 4	c_3, c_{10}	c_1, c_2, c_5
3	p_5=b	3, 8	c_1, c_9	c_2, c_3, c_5
4	p_5=c	6	c_2	c_1, c_{10}
5	p_1=b	7, 10	c_2, c_4	c_3, c_5, c_9
6	p_2=a	9, 11, 14	c_3, c_8	c_4, c_5, c_9
7	p_3=a	12	c_9	c_4, c_5
8	p_4=b	13, 16	c_5, c_6	c_4, c_8, c_9
9	p_5=a	15	c_4	c_8, c_9

TABLEAU 4.7 : Génération des cohortes

Étant donné la cohorte N° 8, une *Ct-Recommandation* consiste à proposer le produit p_4 aux clients c_5 et c_6 qui n'ont pas acheté le produit p_4 à hauteur du niveau b.
De plus, la cohorte N° 5 nous entraîne à recommander le produit p_1 au client c_2 qui

n'a réalisé que 4,21 % de son chiffre d'affaires pour p_1 bien qu'il fasse pourtant parti des gros clients (fort chiffre d'affaires global). Ce type de manque à gagner chez un gros client est généralement difficile à détecter sur le terrain par un commercial.
Enfin, les experts métier s'intéresseront davantage aux cohortes recommandant des produits à hauteur de b et/ou de c, c'est-à-dire des moyens et forts chiffres d'affaires, développant ainsi des volumes plus importants de chiffre d'affaires à chaque recommandation réussie.

Les comportements d'achats sous forme de cohortes étant extraits, concentrons-nous sur l'actionnabilité des contre-exemples.

4.4 Actionnabilité des contre-exemples

Considérons maintenant une cohorte composée de une ou plusieurs règles. Avec leur capacité d'achat potentiellement inexploité, les contre-exemples peuvent contribuer théoriquement au développement du chiffre d'affaires des entreprises. Cependant, tous les contre-exemples d'une cohorte ne présentent pas forcément la même réceptivité face à une recommandation. C'est pourquoi notre méthodologie se concentre sur les clients les plus **actionnables**, c'est-à-dire les contre-exemples les plus « proches » des exemples vis-à-vis de leur comportement d'achat et de leurs variables descriptives (âge, sexe, revenu annuel, etc.). Plus un contre-exemple est proche des exemples, plus il est probable qu'il se comporte comme un exemple, c'est-à-dire qu'il développe davantage son chiffre d'affaires pour la conclusion de la cohorte.

4.4.1 Pré-actionnabilité sur les variables d'achats (étape N° 2, figure 4.1)

Pour être proche de l'ensemble des exemples Ct^+ d'une cohorte, un contre-exemple, noté e^-, ne devrait pas présenter un comportement d'achat « extrême » sur l'ensemble des items en prémisse d'une règle de la cohorte Ct.

DÉFINITION 12
Étant donnés une règle r et un client c, la dépense de c sur r est la somme des parts de CA de c sur les produits des prémisses de la règle :

$$depense(c,r) = \sum_{p_Z \in \text{ prémisses de } r} u^{\%}(p_Z, c) \tag{4.9}$$

DÉFINITION 13
On note :

- *M_r la médiane des dépenses des exemples Ct^+ sur la règle r ;*
- *$(Q_{3_r} - Q_{1_r})$ l'écart interquartile des dépenses des exemples Ct^+ sur la règle r.*

Étant donnée une cohorte Ct, un contre-exemple e^- appartenant à Ct^- est pré-actionnable si et seulement si sa dépense n'est pas extrême par rapport aux dépenses des exemples d'au moins une règle r de la cohorte, c'est-à-dire :

$$M_r - (Q_{3_r} - Q_{1_r}) \leq depense(e^-, r) \leq M_r + (Q_{3_r} - Q_{1_r}) \tag{4.10}$$

EXEMPLE 5 Considérons la cohorte N° 5 recommandant le produit p_1 à hauteur de b, c'est-à-dire entre 15,35 % et 30,30 % de part de chiffre d'affaires. La cohorte est composée de deux règles : la règle N° 7 ($p_5 = b \rightarrow p_1 = b$) et la règle N° 10 ($p_2 = a \rightarrow p_1 = b$). Analysons la répartition des parts de chiffre d'affaires des exemples des règles et vérifions que les contre-exemples vérifient l'inéquation 4.10.

Client	Ensemble	$p_5 = b$	$p_2 = a$
c_3	Ct^+	9,10 %	0 %
c_5	Ct^+	38,78 %	14,86 %
c_9	Ct^+	3,70 %	18,52 %
c_2	Ct^-	71,42 %	0 %
c_4	Ct^-	0 %	3,03 %

TABLEAU 4.8 : Analyse des achats sur les prémisses des règles de la cohorte N° 5

Pour la règle N° 7, $M = 23,94$ % et l'écart interquartile est égal à $[9,10; 38,78$ %$]$. Pour la règle N° 10, $M = 16,69$ % et l'écart interquartile est égal à $[14,86; 18,52$ %$]$. Dès lors, les contre-exemples c_2 et c_4 présentent des comportements d'achat extrêmes car leurs dépenses (71,42 % et 3,03 % respectivement) sur les items de la prémisse des règles sont en dehors de l'intervalle définit dans la formule 4.10.

Une fois ce premier filtrage réalisé (pré-actionnabilité) sur le comportement d'achats des contre-exemples, analysons leurs variables descriptives pour juger de leur actionnabilité.

4.4.2 Actionnabilité sur les variables descriptives (étape N° 3, figure 4.1)

Les variables d'achats sur les produits présentent ce que font les clients, de ce fait elles évoluent chaque jour au cours de l'activité des clients. Les variables descriptives sont elles relativement statiques puisqu'elles présentent ce que sont les clients : l'âge, le sexe, le métier, le revenu annuel, etc.
Nous cherchons à mesurer des proximités entre clients exemples et clients contre-exemples dans l'espace des variables descriptives. Pour cela, nous avons besoin d'une mesure de distance qui puisse prendre en compte à la fois des variables descriptives numériques (ensemble de variables noté N) et des variables descriptives catégoriques (ensemble de variables noté I).
Une analyse factorielle permettant d'inclure à la fois des variables numériques et catégoriques en tant qu'éléments actifs d'une même analyse a été proposée par Escofier en 1979 [86] dans le cadre de l'Analyse des Correspondances Multiples. Pagès [23, 210] souligne que cette approche se confondait avec les travaux de Saporta de 1990 [244] autour de l'Analyse en Composantes Principales. L'ensemble de ces points de vue confère une méthode à part entière dotée de plusieurs bonnes propriétés : l'Analyse Factorielle de Données Mixtes (AFDM). Pour mesurer l'actionnabilité

des clients contre-exemples dans un espace à variables mixtes (numériques et catégoriques), nous utilisons la mesure de distance de l'AFDM dans notre méthodologie *CAPRE*. Le principe général est le suivant : un contre-exemple est considéré comme actionnable s'il est suffisamment proche des exemples.
Les variables descriptives des clients sont séparées en deux groupes distincts :

- Les variables descriptives numériques $N = \{N_1, N_2, \dots N_s\}$ centrées et réduites ;
- Un ensemble d'indicatrices [1] $I = \{I_1, I_2, \dots I_t\}$ obtenu en appliquant un codage disjonctif complet puis une opération de réduction sur les variables descriptives catégoriques.

La distance AFDM entre deux clients c_1 et c_2 dans l'espace des variables descriptives du profil est définie comme suit :

$$d^2(c_1, c_2) = \sum_{i=1}^{s} (N_{i,c_1} - N_{i,c_2})^2 + \sum_{j=1}^{t} (I_{j,c_1} - I_{j,c_2})^2 \tag{4.11}$$

EXEMPLE 6 Chaque client $c \in C$ est décrit par un ensemble de variables descriptives numériques N telles que l'âge ou le revenu annuel, et catégoriques I tel que le sexe (cf. tableau 4.9). À partir de ces variables, la matrice des distances entre clients est calculée (cf. tableau 4.10).

	Âge	**Sexe**	**Revenu annuel**
c_1	30	Homme	83 000
c_2	34	Homme	75 000
c_3	55	Femme	25 000
c_4	42	Homme	28 000
c_5	50	Homme	22 000
c_6	55	Femme	39 000
c_7	25	Homme	45 000
c_8	22	Femme	40 000
c_9	32	Femme	35 000
c_{10}	40	Homme	32 000

TABLEAU 4.9 : Variables descriptives numériques et catégorique

Les trois clients c_1, c_2 et c_3 achètent les trois produits p_1, p_4 et p_5, mais le client c_3 semble éloigné des clients c_1 et c_2 par ses variables descriptives. En effet, les clients c_1 et c_2, hommes d'une trentaine d'année à fort revenu annuel, auraient tendance à acheter les trois produits alors que le client c_3, femme plus âgée et à plus faible revenu, semble moins intéressée par le produit p_3. Du fait de la distance entre clients,

1. Les valeurs de ces indicatrices sont divisées par la racine carrée de la fréquence de chaque indicatrice.

le client c_3, s'il est contre-exemple, sera éloigné des clients c_1 et c_2 (distances respectives de 8,333 et 6,300) si ces derniers sont exemples d'une cohorte. En revanche, les clients c_1 et c_2 sont très proches dans l'espace des variables descriptives (distance de 0,151).

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
c_1	0,000									
c_2	**0,151**	0,000								
c_3	**8,333**	**6,300**	0,000							
c_4	7,119	5,198	0,438	0,000						
c_5	8,757	6,610	0,438	0,085	0,000					
c_6	4,973	3,467	0,461	0,701	1,097	0,000				
c_7	3,398	2,118	1,358	0,680	1,245	0,501	0,000			
c_8	4,768	3,299	0,529	0,756	1,179	0,002	0,475	0,000		
c_9	5,839	4,182	0,235	0,532	0,814	0,038	0,652	0,059	0,000	
c_{10}	6,121	4,351	0,532	0,038	0,235	0,532	0,398	0,567	0,438	0,000

TABLEAU 4.10 : Matrice des similarités clients utilisant la distance AFDM

DÉFINITION 14

*Un contre-exemple e^- est **actionnable** si il est suffisamment proche des exemples. Étant donnés deux seuils minDist et δ, e^- est considéré comme δ-**actionnable** vis-à-vis d'une cohorte Ct si et seulement si :*

1. *e^- est pré-actionnable (cf. section 4.4.1).*
2. *e^- respecte l'inéquation suivante :*

$$\frac{\left|\{e^+ \in Ct^+ |\ d(e^-,\ e^+) \leqslant minDist\}\right|}{\left|Ct^+\right|} \geqslant \delta \tag{4.12}$$

avec minDist, un seuil déterminant un voisinage autour du contre-exemple e^-, et δ un seuil indiquant la proportion minimale d'exemples dans son voisinage pour être actionnable (cf. figure 4.3).

Compte tenu de la distribution des exemples dans l'espace des variables descriptives des clients, la distance d'un contre-exemple au barycentre des exemples n'est pas forcément représentative. C'est pourquoi, nous n'utilisons pas le barycentre dans notre définition de l'actionnabilité.

EXEMPLE 7 Reprenons notre ensemble de cohortes, le paramètre *minDist* correspond à la distance moyenne entre l'ensemble des clients, c'est-à-dire $minDist = 2,47$. Nous fixons empiriquement le paramètre $\delta = 10$ %, c'est-à-dire que chaque contre-exemple pour être actionnable devra être proche d'*au moins 10 %* des exemples à une distance inférieure ou égale à 2,47. Dès lors, le filtrage des contre-exemples (CE) après application des étapes de pré-actionnabilité et d'actionnabilité est résumé dans le tableau 4.11.

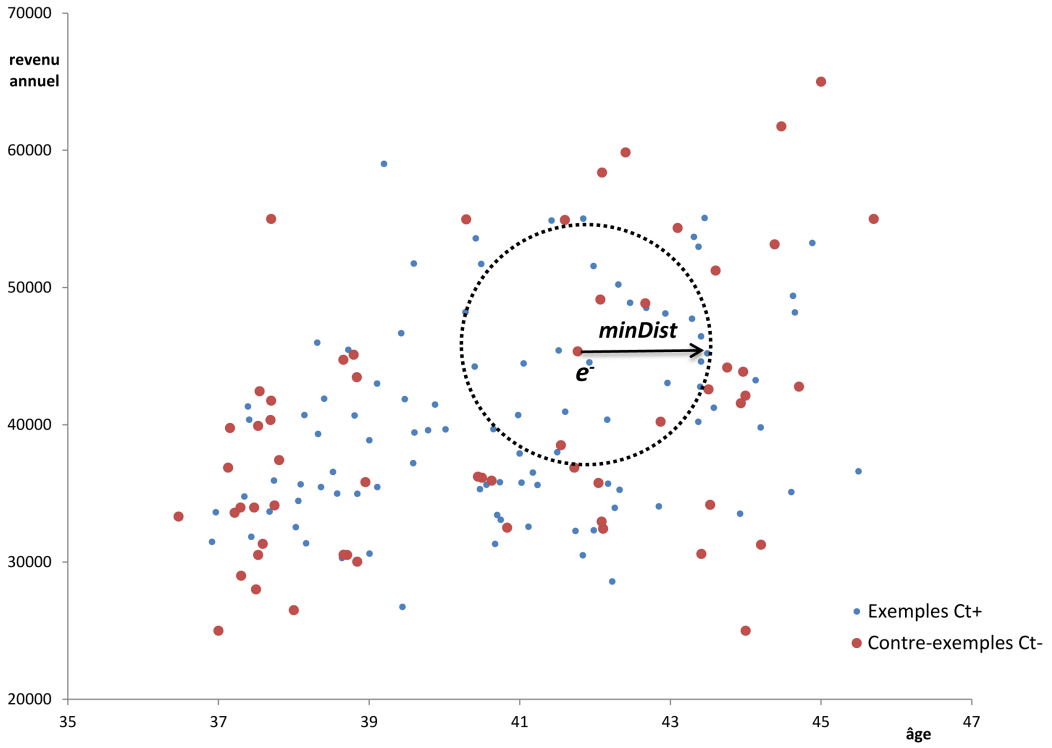


FIGURE 4.3 : Nuage de points des clients dans un espace 2D des variables descriptives

Nº Cohorte	Conclusion	CE	CE pré-actionnables	CE actionnables
1	p_1=a	c_5	c_5	c_5
2	p_4=a	c_3, c_{10}	c_{10}	c_{10}
3	p_5=b	c_1, c_9	c_1, c_9	c_9
4	p_5=c	c_2	c_2	c_2
5	p_1=b	c_2, c_4		
6	p_2=a	c_3, c_8		
7	p_3=a	c_9	c_9	c_9
8	p_4=b	c_5, c_6	c_5, c_6	c_5, c_6
9	p_5=a	c_4		

TABLEAU 4.11 : Filtrage des contre-exemples des cohortes

4.5 Intérêt économique des cohortes (étape Nº 4, figure 4.1)

4.5.1 Profitabilité a priori

L'intérêt économique d'un contre-exemple correspond à la rentabilité en Euros de la Ct-Recommandation appliquée. Pour chaque contre-exemple actionnable (cf. définition 14), nous calculons la somme de chiffre d'affaires en Euros qui aurait due être réalisée si le contre-exemple s'était comporté comme un exemple de la cohorte. C'est un raisonnement optimiste mais qui a l'intérêt de permettre de trier les recommandations et/ou les clients par profit.

DÉFINITION 15

Nous notons $u(\blacksquare, c)$ la somme $\sum_{p \in P} u(p, c)$. Étant donné un seuil $\theta \in \mathbb{R}^+$, un contre-exemple e^- est défini comme θ-profitable pour le produit p_Z, c'est-à-dire pour la recommandation $Ct(p_Z)$ si et seulement si :

$$Profit(Ct(p_Z), e^-) = (u(\blacksquare, e^-) * \alpha) - u(p_Z, e^-) \geqslant \theta$$

avec la part moyenne des dépenses consacrées à p_Z par les exemples : $\alpha = \dfrac{\sum_{c \in Ct^+} u(p_Z, c)}{\sum_{c \in Ct^+} u(\blacksquare, c)}$ (4.13)

EXEMPLE 8 Intéressons-nous aux contre-exemples actionnables 1000 €-profitables des cohortes ($\theta = 1000$ €). Nous pouvons déduire que la rentabilité d'une cohorte de règles correspond à la somme des rentabilités des contre-exemples 1000 €-profitables de la cohorte. Dans le tableau 4.12, seuls les clients c_{10}, c_9 et c_6 sont 1000 €-profitables pour leurs cohortes respectives. Dès lors, le profit a priori global peut être estimé à 13 530,66 € pour les trois recommandations et la recommandation correspondant à la cohorte N° 3 s'avère être la plus profitable avec un profit espéré de 6 069,35 €.

N° Cohorte	CE actionnables	CE 1000 €-profitables	Profit a priori
1	c_5		
2	c_{10}	c_{10}	3411,06 €
3	c_9	c_9	6069,35 €
4	c_2		
5			
6			
7	c_9		
8	c_5, c_6	c_6	4050,25 €
9			

TABLEAU 4.12 : Profit a priori espéré des contre-exemples profitables des cohortes

De cet intérêt économique, pourront être déduits les coûts fixes et variables engendrés par une démarche commerciale.

4.5.2 Profitabilité personnalisée

Pour prédire plus précisément le chiffre d'affaires qu'un contre-exemple aurait dû dépenser sur un item s'il s'était comporté comme un exemple, nous suggérons l'utilisation d'un *scoring* affinant l'intérêt économique espéré pour chaque client [223]. La phase d'apprentissage est guidée par les exemples Ct^+ de la cohorte et la phase d'application par les contre-exemples actionnables. La cible du score correspond à la valeur en Euros de la conclusion de la cohorte. Le résultat du *scoring* correspond au profit espéré, i.e à $u(\blacksquare, e^-) * \alpha$ de l'équation 4.13. Par la suite, nous soustrayons la valeur réellement dépensée par le client pour obtenir le profit personnalisé espéré (chiffre d'affaires additionnel) pour l'item. Par exemple, pour chaque cohorte, nous réalisons un *scoring* avec l'outil de fouille de données KXEN. Les fondements de l'outil sont présentés en annexe B.

4.6 Présentation des cohortes les plus actionnables et profitables (étape N° 5, figure 4.1)

4.6.1 Choix des canaux de communication

Les canaux de communication marketing traditionnels tels que la visite d'un commercial, le téléphone ou le courrier restent fortement utilisés malgré l'émergence de nouveaux médias [17]. L'utilisation de canaux trop diversifiés ou inadaptés, peut en outre avoir un effet néfaste sur la relation avec le client. Quels sont les canaux de communication les plus efficaces pour transformer la recommandation en profitabilité pour l'entreprise ? Il est à première vue difficile d'apporter une réponse précise. Dans la méthodologie *CAPRE*, notre objectif est d'optimiser l'utilisation des canaux de communication. Citons les canaux essentiels dans le démarchage client [160] :

1. **Visite commerciale** : une visite « sur le terrain » d'un commercial proposant au client démarché l'achat d'un ou plusieurs produits recommandés ;
2. **Courrier** : un courrier envoyé directement au client dans sa boîte aux lettres suggérant quelques recommandations et illustrant quelques promotions ;
3. **E-mail** : un e-mail propose quelques recommandations produits et est directement envoyé au client dans sa boîte emails via Internet et des retours statistiques peuvent être collectés sur les préférences des clients ;
4. **Téléphone** : un argumentaire est préparé et utilisé durant les appels téléphoniques aux clients. Nous imaginons que cet argumentaire est relatif aux recommandations ;
5. **SMS** : un SMS ou un MMS suggère quelques promotions d'achats sur les recommandations du système et est envoyé par un fournisseur externe afin de collecter des retours statistiques ;
6. **Fax** : un fax propose quelques recommandations d'achats et est envoyé directement aux clients.

Les caractéristiques des canaux de communication sont abordées dans les parties suivantes.

Coûts fixes et variables

Nous distinguons les coûts fixes (cf. tableau 4.13) et les coûts variables (cf. tableau 4.14) de chaque canal de communication marketing. Ces coûts doivent être déterminés par les experts métier et notamment le directeur marketing. Par exemple, la création d'un email (coûts fixes) est beaucoup plus onéreuse que son envoi (coûts variables), ce dernier variant suivant le nombre de clients n ciblés. Par conséquent, nous pouvons généraliser le coût d'utilisation d'un canal marketing j comme suit :

$$Cout(j) = FC_j + (VC_j * n_j) \tag{4.14}$$

avec, FC_j les coûts fixes du canal j, VC_j les coûts variables par client pour le canal j et n_j le nombre de clients contactés à travers le canal j.

Canal j	**Coûts fixes** FC_j	**Explication**
Visite commerciale	0 €	indépendant de la campagne
Courrier	1 500 €	conception, édition et personnalisation
E-mail	1 000 €	conception et personnalisation
Téléphone	1 500 €	argumentaire et *coaching*
SMS	$\simeq$ 0 €	développement interne à l'entreprise
Fax	1 000 €	conception et personnalisation

TABLEAU 4.13 : Coûts fixes des canaux de communication marketing

Canal j	**Coûts variables** VC_j	**Explication**
Visite commerciale	250 €	salaire, voiture, primes
Courrier	0,81 €	impression et affranchissement
E-mail	0,01 €	sous traitance société externe
Téléphone	3 €	charges de télécommunications
SMS	0,08 €	sous traitance société externe
Fax	0,028 €	sous traitance société externe

TABLEAU 4.14 : Coûts variables des canaux de communication marketing

Critères des canaux

En considérant que les informations relatives au démarchage des clients sont complètes et à jour dans la base de données de l'entreprise, nous mesurons trois critères différents pour estimer le taux de réponse du client pour chaque canal de communication j (cf. tableau 4.15).

- **Capacité** Ca_j : la proportion maximale de clients pouvant être contactés par le canal j en utilisant l'ensemble des ressources et données existantes ;
- **Atteinte** At_j : la proportion de clients effectivement contactés en utilisant le canal de communication j ;
- **Conviction** Cv_j : la proportion de clients convaincus de l'intérêt de la recommandation et ayant été préalablement contactés à travers le canal j.

Ces trois indicateurs doivent également être renseignés par les experts métier de l'entreprise car ils peuvent fluctuer en fonction de la typologie de clients contactés et du domaine de l'entreprise. Dans la notion *in-depth mining* proposée par Cao dans la méthodologie DDID-PD (cf. section 2.2.2.2), une grande attention doit être accordée aux besoins de l'entreprise, à la connaissance du domaine et aux renseignements qualitatifs des experts métier pour impacter directement le processus de fouille de données.

Canal j	**Capacité** Ca_j	**Atteinte** At_j	**Conviction** Cv_j
Visite commerciale	25 %	100 %	80 %
Courrier	100 %	80 %	50 %
E-mail	100 %	30 %	10 %
Téléphone	50 %	60 %	30 %
SMS	100 %	40 %	15 %
Fax	100 %	50 %	20 %

TABLEAU 4.15 : Critères des canaux de communication : capacité, atteinte et conviction

Classes de canaux

Soit une classe Cl de canaux de communication, c'est-à-dire un ensemble compatible de différents canaux de communication marketing : {Visite commerciale, courrier, SMS} par exemple. Réaliser du multi-canal peut améliorer la probabilité d'atteindre et de convaincre le client, il n'est cependant pas conseillé de « sur-communiquer » avec le client.
Dans l'objectif de choisir la meilleure classe de canaux, nous essayons d'optimiser la profitabilité estimée par client en fonction des items recommandés. Si la classe de canaux Cl est utilisée, une estimation de la profitabilité par client i est la suivante :

$$ProfitClient(i, Cl) = p_i * g_i * \max_{j \in Cl}(At_j * Cv_j) - \sum_{j \in Cl} VC_j \tag{4.15}$$

où pour chaque client i la variable p_i correspond à la probabilité de participation à la campagne et g_i le gain espéré de marge nette. Le maximum $\max_{j \in Cl}(At_j * Cv_j)$ est un modèle simplifié puisque nous ne prenons pas en compte les interactions qui peuvent exister entre les différents canaux de communication marketing sur le même client. Ensuite, pour chaque client i, une classe de canaux marketing C_i est choisie maximisant l'équation 4.15 :

$$C_i = arg \max_{Cl} \left(ProfitClient(i, Cl)\right) \tag{4.16}$$

Si l'entreprise souhaite utiliser l'ensemble des canaux de communication, les coûts fixes FC_j seront pour elle une dépense.
Pour résumer, pour chaque canal de communication, les coûts fixes, les coûts variables, la capacité, la probabilité d'atteindre le client et la probabilité de le convaincre doivent être déterminés par les experts métier, ceci pour optimiser l'utilisation des canaux de communication en fonction des items recommandés par le système.

4.6.2 Retour sur investissement

La méthodologie *CAPRE* permet de définir une estimation du ROI espéré avant de déclencher les cohortes les plus actionnables et profitables pour l'entreprise, ceci en supposant que pour chaque client c θ-profitable, la profitabilité $Profit(Ct(p), c)$ est calculée pour les cohortes (cf. équation 4.13) les plus profitables.

$$ROI = \sum_{\substack{c \in C \\ p \in P}} Profit(Ct(p), c) \; - \; \sum_{j} FC_j \; - \; \omega \quad \text{with } \omega = \alpha + \beta - \gamma \tag{4.17}$$

- α : coûts fixes de l'opération : communication, publicité, cadeaux, etc.
- β : coûts de la fouille de données : temps de pré-traitement, modélisation, licence logiciel, etc.
- γ : gain de temps : temps humain non dépensé par l'équipe force de vente à établir des recommandations.

Les coûts moyens de fouille de données dépendent du nombre de modèles industrialisés avec notre méthodologie. A contrario, les coûts de recommandations d'un nouveau client seront marginaux. Les coûts moyens diminuent lorsque le coût marginal est inférieur au coût moyen. Cet exemple illustre le concept de mise à l'échelle (*scale-up*) et souligne l'intérêt d'industrialiser les modèles de fouille de données pour réduire le coût moyen [127].

4.7 Prise de décision des experts métier (étape N° 6, figure 4.1)

Plusieurs scénarii s'offrent aux experts métier pour actionner le système de recommandation de manière économiquement intéressante :

1. Trier les cohortes pour identifier les Ct-Recommandations les plus intéressantes économiquement. Dès lors, des actions commerciales peuvent être déclenchées pour promouvoir des produits.
2. Trier pour chaque client contre-exemple l'ensemble des cohortes dans lesquelles il est θ-profitable. Les commerciaux pourront réaliser une ou plusieurs recommandations, créant ainsi une relation personnalisée avec le client ciblé.

4.8 Synthèse des paramètres de la méthodologie

Le figure 4.4 synthétise l'ensemble des paramètres utilisés dans *CAPRE*.

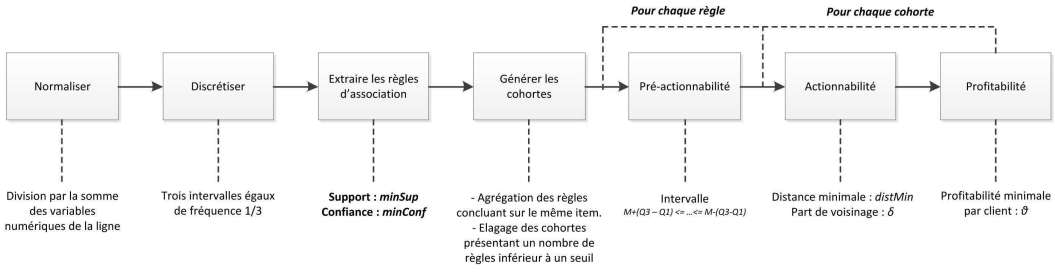


FIGURE 4.4 : Les paramètres de la méthodologie *CAPRE*

La normalisation et la discrétisation dépendent fortement des contraintes du contexte applicatif et notamment de la typologie de clients étudiés. L'extraction des règles est pilotée par deux seuils qui sont le support ($minSup$) et la confiance ($minConf$) permettant aux experts métier de contrôler le nombre de profils d'achats extraits. La génération des cohortes est régi par le nombre d'items différents en conclusion sur l'ensemble des règles ainsi qu'un seuil minimum élaguant les cohortes présentant un nombre de règles trop faible. La phase de pré-actionnabilité est guidée par la médiane et l'intervalle autour duquel les achats extrêmes sont élagués. Cet intervalle peut être pondéré par un coefficient, apportant ainsi un relâchement ou un redressement de la contrainte sur les achats des clients. La phase d'actionnabilité est pilotée par deux paramètres : la distance minimale *minDist* correspond à la moyenne des distances sur l'ensemble de la matrice des clients et la part de voisinage δ est fixée empiriquement par les experts métier. Les paramètres des phases de pré-actionnabilité et d'actionnabilité sont guidés par la prise de risque que souhaitent prendre les experts métier, c'est-à-dire le compromis entre le nombre de fausses recommandations (précision) et le nombre de clients à démarcher (rappel). La phase de profitabilité permet de fixer la rentabilité minimale en Euros devant rapporter chaque client avant de mettre en place une démarche commerciale. Ce seuil ainsi que l'estimation du retour sur investissement dépendent fortement de la stratégie commerciale que souhaite mettre en œuvre l'entreprise.

4.9 Apports de la méthodologie *CAPRE*

La méthodologie CAPRE consiste à extraire des comportements de référence sous la forme de cohortes de règles d'association et à en évaluer l'actionnabilité et l'intérêt économique. Les cohortes concentrent les profils d'achats d'un même produit : regrouper les règles nous fait gagner en compacité de comportements et le regroupement des exemples et des contre-exemples améliore la robustesse des recommandations. Les recommandations sont réalisées en ciblant les contre-exemples les plus actionnables sur les règles les plus rentables. La phase d'actionnabilité de *CAPRE* permet d'identifier les clients présentant un manque à gagner pour lesquels faire une recommandation est objectivement raisonnable (minimisation des fausses recommandations). Nous nous intéressons aux règles d'association non pas seulement en tant qu'implication révélatrice d'un comportement d'achat, mais davantage comme modèle détecteur de contre-exemples à prospecter pour les commerciaux. La profitabilité permet de quantifier le manque à gagner et ainsi d'allouer des moyens en proportion, c'est-à-dire les canaux de communication marketing. Peu de travaux présentent et valident des mesures d'actionnabilité à composante économique ou financière. Dès lors, l'*actionnabilité* et la *profitabilité* constituent une originalité forte de notre méthodologie. Les cohortes présentées aux experts métier représentent un système de recommandations explicites, qualifié de *boîte blanche*. En effet, les commerciaux disposent de règles pour comprendre l'origine de la recommandation et peuvent analyser la population de clients exemples dont les comportements ont amené à l'apparition de la cohorte dans les données.

4.10 Conclusion

Dans ce chapitre, nous avons présenté la méthodologie *CAPRE* (*Customer Actionability and Profitability Recommendation*) pour des recommandations actionnables et profitables de produits à partir de chiffres d'affaires. Pour cela, la méthodologie se décompose en sept étapes : (i) la préparation des données, (ii) l'extraction de règles d'association, (iii) la pré-actionnabilité, (iv) l'actionnabilité, (v) la profitabilité, (vi) le déclenchement des recommandations actionnables et profitables et (vii) la prise de décision des experts métier. Les concepts d'actionnabilité et de profitabilité des recommandations constituent un caractère fort de notre méthodologie en comparaison des autres approches de recommandations. Nous avons également insisté sur l'utilisation des canaux de communication marketing adaptés à la relation avec le client et à l'optimisation du retour sur investissement pour les experts métier. Dans le chapitre 5, nous présentons une implémentation de la méthodologie *CAPRE* nommée *ARKIS* et nous évaluons l'efficacité de notre système sur le jeu de données de référence *MovieLens*.

5

Développements, validation et discussion

A successful tool is one that was used to do something undreamed of by its author [...]

Stephen C. Johnson, 2008

SOMMAIRE

5.1 *ARKIS* : un outil de recommandations actionnables et profitables

La méthodologie *CAPRE* présentée dans le chapitre 4 définit les principes de base pour l'élaboration d'un outil de recommandations actionnables et profitables. La méthodologie peut cependant être implémentée de multiples façons. En particulier, diverses possibilités sont envisageables pour l'extraction de règles d'association. Dans ce chapitre, nous décrivons les choix qui ont été effectués pour mettre en œuvre la méthodologie *CAPRE* dans l'outil *ARKIS* (*Association Rule Knowledge Interactive Search*). Dans la section 5.1, nous présentons les choix d'architecture, d'implémentation et de développement de l'outil. Ensuite, nous détaillons un exemple d'utilisation d'*ARKIS* au travers d'un jeu de données de taille réduite. Enfin, nous présentons une validation des résultats de la méthodologie à travers une expérimentation sur le jeu de données *MovieLens* aboutissant à une comparaison et une discussion des résultats (cf. section 5.2.4).

5.1.1 Architecture

Dans cette section nous présentons l'approche générale de l'outil *ARKIS*, respectant les différentes étapes de la méthodologie présentée dans le figure 4.1.

5.1.1.1 Architecture générale

Deux processus sont implémentés dans *ARKIS* (cf. figure 5.1). Tout d'abord, à partir du chargement des variables d'achats *(1)* (matrice clients $\times$ produits), les données sont préparées pour l'extraction des règles et la génération des cohortes. Dès lors, les phases de pré-actionnabilité *(2)*, d'actionnabilité *(3)* et de profitabilité *(4)* sont déclenchées. Ensuite, la phase d'actionnabilité nécessite le chargement des variables descriptives *(5)* et la génération de la matrice des distances entre clients.

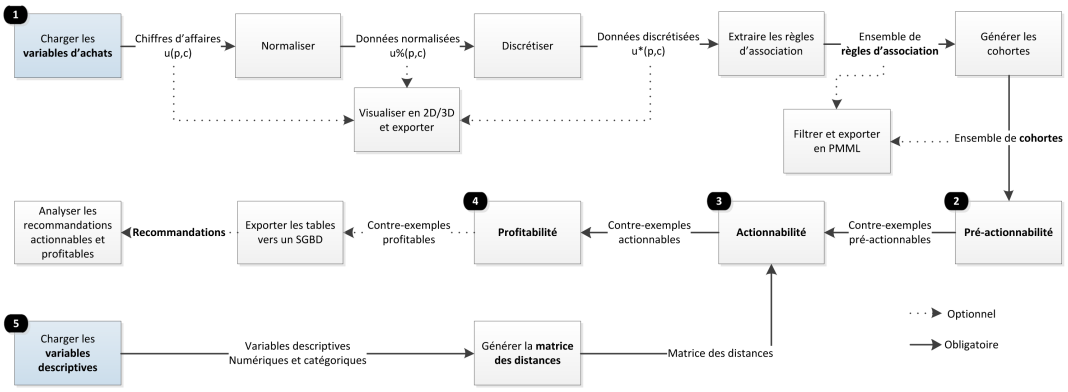


FIGURE 5.1 : Architecture générale de l'implémentation de *CAPRE* dans *ARKIS*

5.1.1.2 Interopérabilité

L'outil *ARKIS* se veut interopérable avec un Système de Gestion de Bases de Données ou SGBD (cf. figure 5.2). En effet, une requête ou procédure stockée dans le SGBD suffit pour créer les deux fichiers CSV [1] des variables d'achats et des variables descriptives (cf. *(1)* et *(5)* figure 5.1). De plus, l'outil *ARKIS* prévoit l'export des modèles au format PMML (cf. section 5.1.2.2) de manière à être compatible avec d'autres outils de fouille de données tels que *ARIPSO* [2] ou *ARVAL* [3]. De la même manière, l'export de toutes les données est réalisé au format CSV, permettant ainsi une réintégration des connaissances dans le SGBD pour une consolidation sous forme de rapport dynamique ou de cube multidimensionnel.

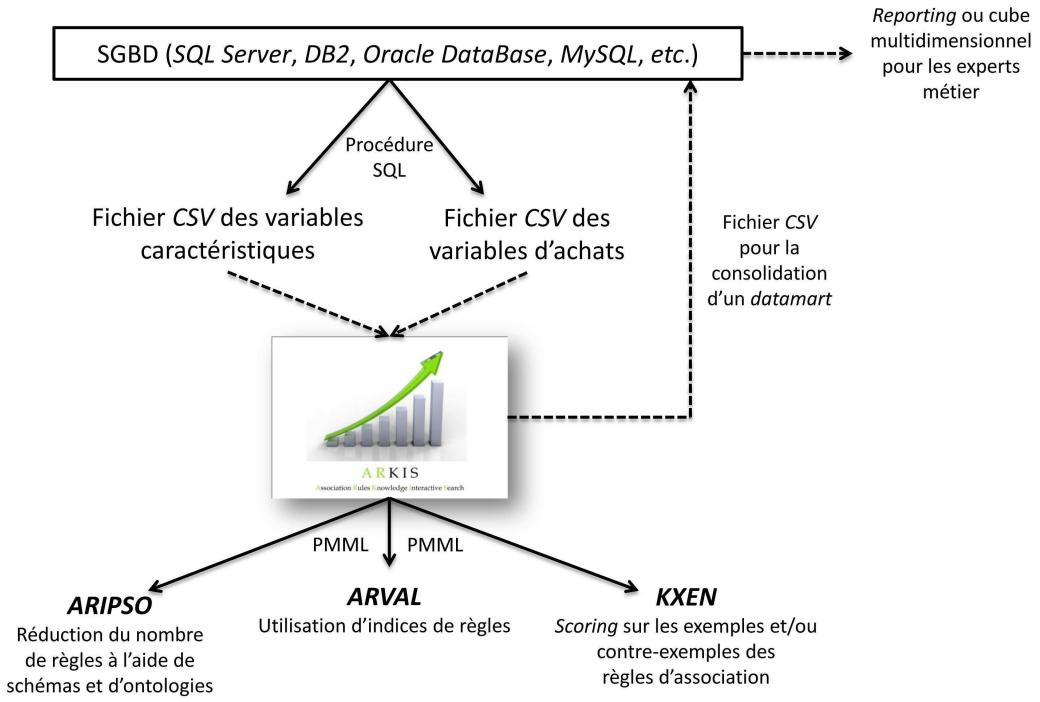


FIGURE 5.2 : Interopérabilité d'*ARKIS* avec un SGBD et des outils de fouille de données

1. *Comma-Separated Values*
2. *ARIPSO* (*Association Rule Interactive Post-processing using Schemas and Ontologies*) est un logiciel libre, développé à l'université de Nantes, permettant à l'utilisateur de réduire le volume de règles découvertes et ainsi d'améliorer la qualité des connaissances extraites.
3. *ARVAL* (*Association Rule's VALidation*) est un logiciel libre, développé à l'université de Nantes, permettant d'appliquer des mesures objectives à des items, des itemsets et des règles d'association. Il accepte les formats WEKA, CSV et PMML en entrée et est disponible à l'adresse http://www.polytech.univ-nantes.fr/arval.

5.1.2 Implémentation

5.1.2.1 Langage, plateforme et librairies

L'outil *ARKIS* a été développé en Java (*Java SE 6*) sous l'environnement de développement *Eclipse* (*Eclipse Platform Version 3.4.2 (Ganymède)*). Le programme d'extraction des règles d'association *CHARM* (cf. section 5.1.2.3) développé en C++ par son concepteur Mohammed J. ZAKI[4] a été compilé dans le cadre du projet avec l'environnement de développement libre *MinGW32*. L'appel de *CHARM* est effectué par des processus existant aussi bien sous Windows que sous Unix, facilitant ainsi la portabilité.
Le développement s'appuie sur un certain nombre de classes et de *packages* pouvant être organisés selon les modules fonctionnels suivants :

- **Gestion de projet** : créer, ouvrir , enregistrer et exporter des projets *ARKIS* ;
- **Préparation des données** : visualiser, (dé)-normaliser, sélectionner des variables, discrétiser (intervalles égaux, fréquences égales ou paramétrables) et exporter des données au format CSV ;
- **Génération de la matrice des distances** : charger le fichier des variables descriptives et générer la matrice des distances ;
- **Extraction des règles** : exécuter l'algorithme d'extraction des règles d'association à partir de deux paramètres *minSup* et *minConf*, filtrer les règles à l'aide de mesures d'intérêts (support, confiance, *lift*, IPEE, intensité implication et taux informationnel) et exporter au format PMML ;
- **Génération des cohortes** : générer les cohortes de règles d'association, former les ensembles d'exemples et de contre-exemples, filtrer les cohortes à l'aide de mesures (nombre de règles, support, *lift* moyen pondéré) et exporter au format CSV ;
- **Pré-actionnabilité** : exécuter la pré-actionnabilité sur les contre-exemples de la cohorte en fonction des deux paramètres *minDist* (distance moyenne de la matrice générée) et δ (part de voisinage) générant la liste des contre-exemples pré-actionnables ;
- **Actionnabilité** : exécuter l'actionnabilité sur les contre-exemples pré-actionnables utilisant la matrice des distances sur les variables descriptives et générant la liste des contre-exemples actionnables ;
- **Profitabilité** : exécuter la profitabilité en fonction du paramètre θ générant ainsi la liste des contre-exemples θ-profitables.

Les caractéristiques globales du développement d'*ARKIS* sont présentées dans le tableau 5.1.

4. http://www.cs.rpi.edu/~zaki/software

Nombre de packages	16
Nombre de classes	109
Nombre de méthodes	605
Nombre de lignes de codes	13 069
Nombre de procédures SQL	8
Charge de développement	120 j/h

TABLEAU 5.1 : Caractéristiques de développement de l'application *ARKIS*

Concernant la mise en place de l'interface graphique, nous utilisons *Visual Editor* (VE) [5]. VE est une extension du projet *Eclipse* offrant les mêmes fonctionnalités que *Jigloo* [6]. En revanche, VE offre une meilleure cohérence du code généré, en particulier lors des retours arrière, et bénéficie du support de la communauté *Eclipse*. Les statistiques sont affichées sous forme graphique avec l'outil *JfreeChart* [7], bibliothèque libre permettant d'afficher différents types de graphes. Enfin, l'affichage des données en trois dimensions a été réalisé avec la librairie *Java3D* [8]. Dans les parties suivantes, nous décrivons succinctement *Swing*, *Java3D* et *JFreeChart*.

Swing est une librairie de génération d'interfaces graphiques faisant partie du framework *Java J2SE*. Swing utilise les concepts de l'architecture MVC qui signifie « Modèle Vue Contrôleur » (cf. figure 5.3) et permet de développer l'interface utilisateur indépendamment de la plateforme matérielle ou logicielle utilisée.

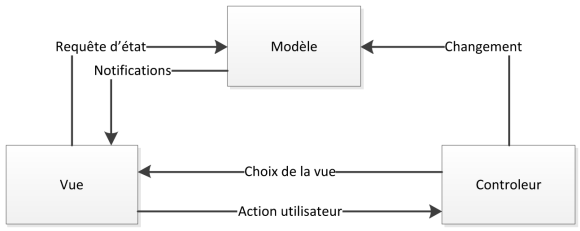


FIGURE 5.3 : Architecture Modèle Vue Contrôleur

- Le **modèle** représente les données de l'application (et les interactions avec la base de données), c'est-à-dire l'*état* de l'application. Le modèle ne connaît rien de ses contrôleurs ou de ses vues ;
- La **vue** correspond à la représentation visuelle des données dans l'état actuel que lui fournit le modèle ;
- Le **contrôleur** gère l'interaction utilisateur avec le modèle. Il intercepte les entrées utilisateurs dans la vue et les traduit en changeant l'état du modèle.

5. http://www.eclipse.org/vep
6. http://www.cloudgarden.com/jigloo
7. http://www.jfree.org/jfreechart
8. https://java3d.dev.java.net

Swing supporte l'architecture MVC explicitement avec des *widgets* tels que *JList* ou *JTree*. Les objets graphiques Swing sont décomposés en deux classes : les composants (*components*), éléments de base (par exemple un bouton) et les contenants (*containers*, par exemple une fenêtre de dialogue) contenant eux mêmes les composants. Chaque objet comporte des attributs définissant ses caractéristiques graphiques (taille, texte, couleur, etc.) et des méthodes. Une interface utilisateur générée avec Swing est formée d'une arborescence d'objets graphiques dont la racine est un contenant et les nœuds des composants ou des contenants.

Java3D est une librairie de visualisation en trois dimensions développée par *Sun* qui offre les outils nécessaires pour la génération d'objets complexes, le rendu, l'éclairage et la navigation dans l'univers créé. Une scène Java3D est composée d'un graphe acyclique chaînant des objets qui définissent la représentation géométrique des éléments affichés, leurs déplacements et la gestion des points de vue [9]. La construction d'un univers Java3D suit les étapes suivantes :

1. Construction de l'univers :
   (a) Création du *Canvas3D* ;
   (b) Création de l'*Universe* et rattachement au *Canvas3D* ;
   (c) Création du *BranchGroup* et rattachement à l'*Universe* ;
2. Construction de chaque objet à intégrer dans l'univers :
   (a) Création d'un *ObjTransformGroup* et rattachement au *BranchGroup* ;
   (b) Création d'un *TransformGroup* et rattachement à l'*ObjTransformGroup* ;
   (c) Création de l'objet et rattachement au *TransformGroup*.

L'API Java3D se base sur des technologies déjà existantes telles que *DirectX* et *OpenGL*. Java3D peut également implémenter des objets créés avec des programmes de modélisation 3D tels que *TrueSpace* ou *VRML*.

JFreeChart [10] est une librairie de visualisation de données sous forme de graphes. Partant d'une structure de données codées sous la forme d'une suite de couples (libellé, valeur). Cette librairie permet la génération d'un objet Swing contenant la représentation de ces données sous forme de diagrammes en secteurs ou d'histogrammes par exemple. L'affichage des données comporte trois étapes :

1. L'instanciation et l'initialisation de la classe *JFreeChart* qui contient la définition de l'objet graphique ;
2. L'instanciation de la classe *DataSet* et le chargement des données à afficher ;
3. L'instanciation de la classe *ChartPanel* générant un *JPanel* Swing contenant le graphique.

9. http://java.sun.com/developer/onlineTraining/java3d/j3d_tutorial_ch1.pdf
10. http://www.jfree.org/jfreechart

5.1.2.2 Outils et normes

L'application *ARKIS* est spécifiée en UML à l'aide du logiciel de modélisation *Visio*. La documentation technique est fournie sous la forme d'une *JavaDoc* [11]. Le code source du logiciel est géré via le logiciel de gestion de versions *SubVersion* [12], facilitant le travail collaboratif entre le laboratoire et l'entreprise.

UML *(Unified Modeling Langage)* ou langage de modélisation unifié constitue un langage permettant d'exprimer des relations, des comportements et des notions de haut niveau sur les objets. UML est issu de la fusion de trois méthodes de modélisation distinctes : la méthode *Booch* développée par Grady Booch ; la technique de modélisation objet (*Object Modeling Technique*, OMT) développée par James Rumbaugh et la méthode Objectory, conçue par Ivar Jacobson. La version 2.0 d'UML se compose de quatre grandes spécifications : la spécification d'échanges de diagrammes (*Diagram Interchange Specification*), l'infrastructure d'UML, la superstructure d'UML et le langage de contraintes d'objets (*Object Constraint Language*, OCL). Toutes ces spécifications sont disponibles sur le site de l'OMG [13].

PMML *(Predictive Modeling Markup Language)* est un langage créé en 1998 par le *Data Mining Group* [14] basé sur le XML. Il définit une manière standard de décrire les modèles statistiques et de traitements des données au sein des applications décisionnelles. L'objectif est de partager des modèles analytiques en dépassant les frontières technologiques des outils du marché. Le PMML a pour vocation de couvrir l'ensemble des méthodes de fouille de données. De nombreux acteurs majeurs du marché y sont fortement impliqués [308] tels que *IBM*, *KXEN*, *MicroStrategy* et *SAP*. Le *Data Mining Group* offre la possibilité aux *data miners* de contrôler la conformité de leur modèle en ligne (cf. figure 5.4).

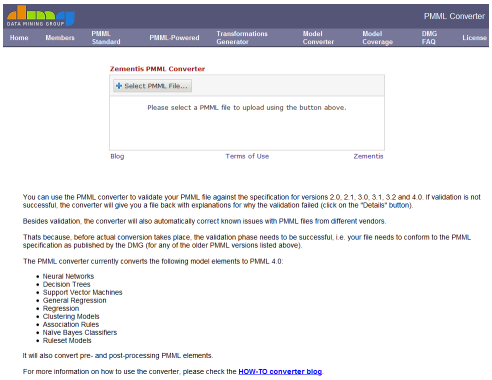


FIGURE 5.4 : Contrôle de la véracité de l'export PMML

11. http://java.sun.com/j2se/javadoc
12. http://subversion.tigris.org
13. http://www.omg.org
14. http://www.dmg.org

5.1.2.3 Choix de l'algorithme d'extraction de règles

Dans cette sous section, nous nous intéressons à l'étape N° 1 de notre méthodologie (cf. chapitre 4).

Présentation des algorithmes
Différents algorithmes d'extraction de règles d'association sont disponibles pour la mise au point d'un outil d'extraction de comportements d'achats sous forme de règles. Un certain nombre s'inspire de l'algorithme *Apriori*. Ces algorithmes ont la particularité de générer un nombre exorbitant de règles, ce qui rend leur exploitation quasiment impossible par des experts métier.
Soient *minSup* et *minConf* les seuils de support et de confiance fixés par l'utilisateur. Les algorithmes d'extraction de règles d'association peuvent être décomposés en deux sous problèmes [32] :

1. Trouver tous les items fréquents apparaissant dans la base de données avec une fréquence supérieure ou égale au seuil *minSup* ;
2. À partir des itemsets fréquents, générer toutes les règles d'association ayant une mesure de confiance supérieure ou égale à *minConf*.

La taille de l'espace de recherche pour l'extraction des itemsets fréquents est exponentielle avec le nombre m d'items. En effet, il y a potentiellement 2^m-1 itemsets fréquents. Ainsi, suite à l'apparition de l'algorithme *Apriori* [5], des travaux ont été réalisés pour trouver les bonnes heuristiques pour l'élagage de l'espace de recherche. De manière générale, nous pouvons dire que les algorithmes de découverte de règles d'association doivent faire face aux défis suivants [299] :

- Trouver des heuristiques syntaxiques (sous-ensembles d'itemsets) ou basées sur des métriques statistiques (inférieur à *minSup*) pour élaguer l'espace de recherche ;
- Trouver des techniques pour réduire les coûts d'entrée/sortie ;
- Trouver des solutions algorithmiques (par exemple des structures de données adéquates) pour minimiser le coût de l'étape de calcul de support des itemsets.

Algorithme d'extraction d'itemsets fréquents : *Apriori*
L'idée de l'algorithme *Apriori* est de réaliser l'extraction des itemsets fréquents niveau par niveau : il parcourt l'espace de recherche en largeur d'abord, retient les itemsets fréquents et génère d'autres éléments dans le niveau suivant par jointure. Les supports des itemsets candidats sont calculés et les candidats non fréquents sont élagués. Cet élagage est basé essentiellement sur la propriété d'anti-monotonie du support : « si un itemset n'est pas fréquent, aucun de ses sur-itemsets n'est fréquent », à savoir que le support est monotone et sa valeur décroît par rapport à l'extension des itemsets.

Pour déterminer l'ensemble F_k des k-itemsets fréquents, l'algorithme construit un ensemble C_k de k-itemsets candidats (c'est-à-dire susceptibles d'être fréquents) à partir de l'ensemble F_{k-1} des (k-1)-itemsets fréquents extraits au niveau précédent. Plus précisément, à chaque niveau k, l'algorithme effectue deux opérations :

1. La génération de C_k à partir de F_{k-1} (lignes 5 à 6 dans l'algorithme 1) ;

2. L'élagage de C_k pour déterminer F_k (lignes 7 à 13 dans l'algorithme 1).

L'algorithme fournit l'ensemble des itemsets fréquents accompagnés de leur support respectif pour calculer les indices de règles [32].

Algorithme 1: Extraction des itemsets fréquents dans *Apriori* [32]

Entrées : • $\mathcal{BD}$ la base de données et $minSup$ le seuil de support minimal

Sorties : • L l'ensemble de couples $(i, support(i))$ où i est un itemset et $support(i)$ son support

1 $F_1 =$ `extraireItemsFréquents(`$\mathcal{BD}, minSup$`)`2; $n =$ `nombreIndividus(`$\mathcal{BD}$`)`3; $k = 2$;
4 **tant que** $F_{k-1} \neq \emptyset$ **faire**
5 `//Génération de` C_k `à partir de` F_{k-1} ;
6 $C_k =$ `générerItemsetsCandidats(`F_{k-1}`)` ;
7 `//Elagage de` C_k `pour déterminer` F_k ;
8 $C_k =$ `élagagePréliminaire(`C_k, F_{k-1}`)` ; **pour chaque** *individu* $\in \mathcal{BD}$ **faire**
10 **pour chaque** $i \in C_k$ **faire**
11 **si** *individu* `vérifie` i **alors**
12 $i.compteur$++ ;
13 $F_k = \left\{ (i, \frac{i.compteur}{n}) \mid i \in C_k \text{ et } \frac{i.compteur}{n} \geq minSup \right\}$; $k++$;
15 $L = \bigcup_k F_k$; **retourne** L ;

À partir d'un itemset fréquent i, l'algorithme construit toutes les règles de la forme $a \rightarrow b$ où a et b sont deux sous-itemsets de i qui ne possèdent pas d'item en commun et qui redonnent i par conjonction : $a \wedge b = i$. La confiance d'une telle règle est calculée de la manière suivante : $confiance\,(a \rightarrow b) = \frac{support(i)}{support(a)}$. Le fait que i soit un itemset fréquent garantit que son sous-itemset a l'est aussi et donc que son support est connu (déterminé à l'étape d'extraction des itemsets fréquents). Les règles retournées par l'algorithme sont celles dont la confiance est supérieure au seuil de confiance minimale $minConf$.

L'inconvénient majeur de cet algorithme est que le nombre d'itemsets fréquents extraits et leur taille moyenne sont élevés. Le nombre de règles d'association générées varie en général de plusieurs dizaines de milliers à plusieurs millions [305]. Le problème de la pertinence et de l'utilité des règles extraites demeure un problème majeur pour l'extraction des règles d'association. Il est lié à la présence d'une forte proportion de règles redondantes. En revanche, les approches basées sur la découverte d'itemsets *fermés*, proposent de ne générer qu'un sous ensemble compact et générique des règles d'association. Ce sous-ensemble présente l'avantage d'être complet d'un point de vue connaissance tout en présentant une taille réduite [18].

EXEMPLE 9 Considérons deux transactions $\{i_1, i_2, ..., i_{50}\}$ et $\{i_1, i_2, ..., i_{100}\}$. L'algorithme *Apriori* calcule les supports de $2^{100} - 1 \simeq 10^{30}$ itemsets pour $minSup = 1$ et $minConf = 0,5$. En revanche, un algorithme d'extraction d'itemsets fermés calcule le support de deux itemsets fermés et génère une seule règle d'association.

Algorithme d'extraction d'itemsets fermés : *CHARM*

L'algorithme *CHARM*, proposé par Zaki et al. [306] en 2002, privilégie une exploration en profondeur d'abord de l'espace de recherche. L'idée est d'exploiter la maximalité d'un itemset fermé : « un itemset fermé couplé avec l'ensemble des objets le vérifiant n'est pas inclus dans un autre itemset fermé ». Ainsi, l'algorithme *CHARM* explore simultanément l'espace de recherche des itemsets et celui des identificateurs des transactions dans une structure appelée *IT-Tree* (*Itemset-Tidset Tree*). L'algorithme utilise une représentation verticale, appelée *diffset*, accélérant ainsi le calcul des supports. Le pseudo-code de l'algorithme *CHARM* est donné par l'algorithme 2 [299].

CHARM commence par initialiser la classe de préfixes [P] des nœuds à examiner par les *1-itemsets* fréquents et leurs *TIDsets* associés. Les deux étapes principales sont instanciées comme suit :

1. **L'étape d'élagage** est implémentée via la procédure *CHARM-Propriete*. Cette procédure peut modifier la classe courante $[P]$ en supprimant des *IT-paires* ou en insérant de nouvelles paires dans $[P_i]$. Une *IT-paire* est d'abord élaguée comparativement à *minSup*. L'algorithme vérifie ensuite si cette *IT-paire* est maximale en regardant si son *TIDSet* est inclus dans celui de la paire l'ayant généré. Une fois toutes les *IT-paires* traitées, la nouvelle classe $[P_i]$ est récursivement explorée en profondeur d'abord, en appelant la procédure *CHARM-Étend*.

2. **L'étape de construction** est implémentée via la procédure *CHARM-Étend*, elle combine les *IT-paires* présents dans la classe des préfixes $[P]$. Chaque *IT-paire* $X_i \times (X_i)'$ est combinée par les autres paires $X_j \times (X_j)'$ la suivant dans l'ordre lexicographique. Chaque X_i va générer une nouvelle classe de préfixe $[P_i]$, qui serait initialement vide. Les deux *IT-paires* combinées vont produire une nouvelle paire $X \times Y$, où $X = X_i \cup X_j$ et $Y = (X_i)' \cap (X_j)'$.

Algorithme 2: Algorithme *CHARM* [299]

Entrées : $\mathcal{BD}$ et *minSup*
Sorties : $\mathcal{FC}$, l'ensemble de itemsets fermés fréquents

$[P] = \{X_i \times (X_i)' : X_i \in I \wedge support(X_i) \geq minSup\}$; *CHARM-Étend*($[P]$,$\mathcal{FC} = \emptyset$) ;
retourne $\mathcal{FC}$;

Algorithme 3: Procédure *CHARM-Étend* [299]

Entrées : $[P]$, $\mathcal{FC}$
Sorties : $\mathcal{FC}$

pour tous les $X_i \times (X_i)' \in [P]$ **faire**
  $[P_i] = \emptyset$ et $X = X_i$; **pour tous les** $X_j \times (X_j)' \in [P] \wedge X_j > X_i$ **faire**
    $X = X \cup X_i$; $Y = (X_i)' \cap (X_j)'$; *CHARM-Propriété*($[P]$,$[P_i]$) ; **si** $[P_i] \neq \emptyset$ **alors**
      CHARM-Étend($[P_i]$,$\mathcal{FC}$) ; *Supprimer*($[P_i]$) ; $\mathcal{FC} = \mathcal{FC} \cup X$;
retourne $\mathcal{FC}$

Algorithme 4: Procédure *CHARM-Propriété* [299]

Entrées : $[P]$, $[P_i]$
Sorties : $[P]$

si $support(X) \geq minSup$ **alors**
  si $(X_i)' = (X_j)'$ **alors**
    Supprimer X_j de $[P]$; Remplacer tout X_i par X
  sinon
    si $(X_i)' \subset (X_j)'$ **alors**
      Remplacer tout X_i par X
    sinon
      si $(X_i)' \supset (X_j)'$ **alors**
        Remplacer X_j de $[P]$ Ajouter $X \times Y$ à $[P_i]$
      sinon
        si $(X_i)' \neq (X_j)'$ **alors**
          Ajouter $X \times Y$ à $[P_i]$
retourne $[P]$

EXEMPLE 10 La figure 5.5 [299] illustre l'exécution du processus de découverte des itemsets fermés pour $minSup = 2$.

- *CHARM* commence par initialiser la classe racine par $[] = \{A \times 135, B \times 2345, C \times 1235, E \times 2345\}$.
- Le *1-itemset* D est élagué puisqu'il est non fréquent (étape 2). Au début, le nœud $A \times 135 (X = A)$ est traité, et sera combiné avec les autres éléments.
- Les *1-itemsets* A et B sont combinés, AB inséré dans [A] puisque(A)' diffère de (B)' (étape 14 de la procédure *CHARM-Propriété*).
- Les *1-itemsets* A et C sont combinés. Comme on a $(A)' \subset (C)'$, alors toutes les occurrences de A sont remplacées par AC (étape 8 de la procédure *CHARM-Propriété*). En combinant AC et E, on trouve que (AC)' est différent de (E)'. Dans ce cas, ACE est ajouté à $[AC]$ (étape 14 de la procédure *CHARM-Propriété*).
- Un appel récursif à la procédure *CHARM-Étend* ayant la classe $[AC]$ comme entrée. Le même processus continue jusqu'à l'insertion de [ABCE] dans l'ensemble des itemsets fermés fréquents.
- La branche B est exploitée. Ainsi B et C sont combinés. Comme (B)' est différent de (C)', alors BC est inséré dans la nouvelle classe de [B]. B et E sont combinés et on se trouve dans le cas où (B)' est égal à (E)'. Ainsi E est enlevé de [] et toutes les occurrences de B sont remplacées par BE (étapes 4-5 de la procédure *CHARM-Propriété*).
- Un appel récursif de *CHARM-Étend*, ayant la classe $[BC]$ comme entrée, est effectué. Comme il y a un seul élément, BCE est inséré dans la liste des itemsets fermés fréquents (étape 11 de la procédure *CHARM-Étend*).
- Pour la branche C et comme elle ne peut pas être étendue, alors C est inséré dans la liste des itemsets fermés fréquents.
- La liste des itemsets fermés fréquents consiste en l'ensemble des *IT-paires* non barrées dans la figure 5.5.

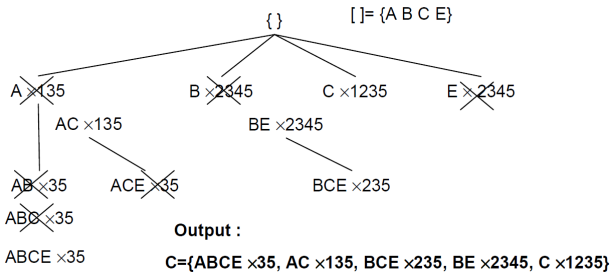


FIGURE 5.5 : Exécution de *CHARM* pour $minSup = 2$ [299]

Comparaison

Comparons les deux algorithmes *Apriori* [15] et *CHARM* [16] sur le jeu de données *Mushroom* [17] (cf. tableau 5.2). *Mushroom* est un jeu de données de référence composé de 8 124 observations (champignons) et 22 attributs tels que la surface, l'odeur, la couleur, etc., ainsi que la cible : comestible ou vénéneux.

Support (%)	**Confiance (%)**	**Nombre de règles *Apriori***	**Nombre de règles *CHARM***
0,9	0,95	13 (cf. tableau 5.3)	7 (cf. tableau 5.4)
0,8	0,95	50	12
0,7	0,95	78	15
0,6	0,95	118	22
0,5	0,95	375	44
0,4	0,95	1 457	103
0,3	0,95	8 499	249

TABLEAU 5.2 : Exécution d'*Apriori* et *CHARM* sur le jeu de données *Mushroom*

La comparaison des règles obtenues permet d'observer qu'*Apriori* extrait de nombreuses règles redondantes dues à la non prise en compte d'itemsets fermés.

Support (%)	**Confiance (%)**	**Règles**
0,97	1	veil-color=w → veil-type=p
0,97	1	gill-attachment=f → veil-type=p
0,97	1	gill-attachment=f & veil-color=w → veil-type=p
0,92	1	ring-number=o → veil-type=p
0,97	0,99	gill-attachment=f → veil-color=w
0,97	1	gill-attachment=f & veil-type=p → veil-color=w
0,97	1	gill-attachment=f → veil-type=p veil-color=w
0,97	0,99	veil-color=w → gill-attachment=f
0,97	1	veil-type=p & veil-color=w → gill-attachment=f
0,97	1	veil-color=w → gill-attachment=f veil-type=p
0,97	0,97	veil-type=p → veil-color=w
0,97	0,97	veil-type=p → gill-attachment=f
0,97	0,97	veil-type=p → gill-attachment=f veil-color=w

TABLEAU 5.3 : Exemples de règles extraites avec *Apriori*, $minSup = 0,9$, $minConf = 0,95$

Par exemple, la règle *gill-attachment=f → veil-type=p* présente un support de 0,97 et une confiance de 1. La règle *gill-attachment=f & veil-color=w → veil-type=p* présente un support de 0,97 et une confiance de 1. Cette deuxième règle est fortement redondante à la première. *Apriori* extrait cette règle, en revanche, *CHARM* l'élague. De plus, le temps d'extraction des règles est beaucoup plus important avec *Apriori* qu'avec *CHARM*. Les performances de l'algorithme *CHARM* ont été évaluées comparativement aux algorithmes *Apriori*, *Close*, *Pascal* et *Closet* dans [19]. Les résultats des expérimentations ont montré que l'algorithme *CHARM* présente de meilleures performances que ses concurrents sur des contextes épars et denses.

15. http://www.cs.waikato.ac.nz/ml/weka
16. http://www.cs.rpi.edu/~zaki/software
17. http://archive.ics.uci.edu/ml/datasets/Mushroom

Support (%)	Confiance (%)	Règles
0,97	1	gill-attachment=f → veil-type=p
0,97	0,97	veil-type=p → gill-attachment=f
0,97	1	veil-color=w → veil-type=p
0,97	0,97	veil-type=p → veil-color=w
0,92	1	ring-number=o → veil-type=p
0,97	0,99	gill-attachment=f → veil-color=w
0,97	0,99	veil-color=w → gill-attachment=f

TABLEAU 5.4 : Exemples de règles extraites avec *CHARM*, $minSup = 0,9$, $minConf = 0,95$

L'inconvénient majeur de l'algorithme *CHARM* est qu'il nécessite un espace important de stockage. En effet, le fait de stocker les itemsets et leurs *TIDsets* accroîts la quantité de mémoire utilisée. Dans la méthodologie *CAPRE*, nous utilisons l'algorithme *CHARM* pour extraire les règles de comportements d'achats. En effet, la réduction de la redondance s'avère un élément différenciateur pour la validation des règles par les experts métier. Dans un contexte industriel, nous privilégions davantage la rapidité d'exécution de l'algorithme que l'espace mémoire requis.

5.1.3 Modélisation

Dans cette sous section, nous nous intéressons à la transition entre les étapes N° 1 et N° 2 de notre méthodologie (cf. chapitre 4).

5.1.3.1 Diagramme de classe simplifié

Le diagramme de classe simplifié de l'application *ARKIS* présenté en figure 5.6 illustre la relation existant entre l'extraction des règles et la génération des cohortes.

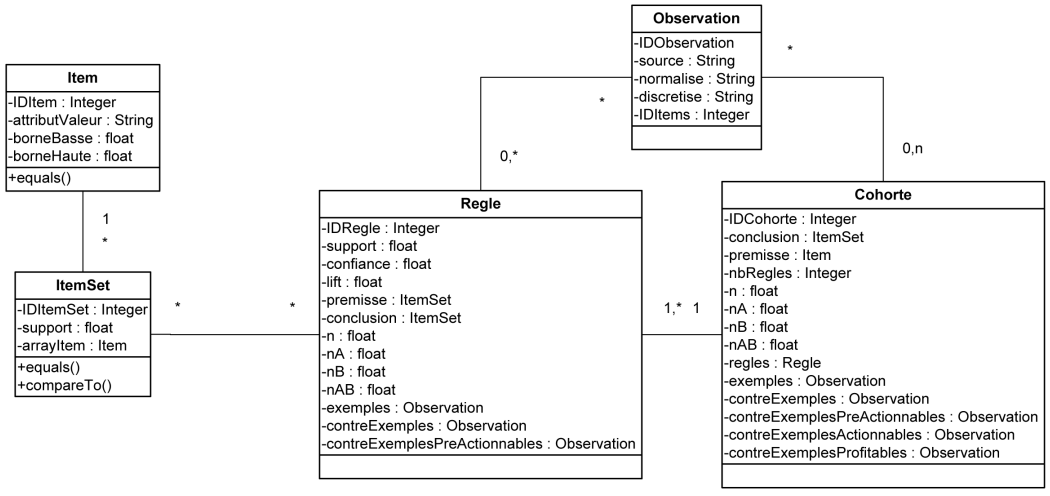


FIGURE 5.6 : Diagramme de classe simplifié d'*ARKIS*

Une règle d'association appartient à une et une seule cohorte. Une cohorte est composée de une ou plusieurs règles d'association. L'extraction des contre-exemples pré-actionnables correspond à l'étape N° 2 de notre méthodologie. Ces contre-exemples pré-actionnables sont stockés au niveau de la classe *Regle* (propriété *contreExemples-PreActionnables*), l'union de ces contre-exemples étant stockée de la même manière dans la classe *Cohorte*. Enfin, la génération des contre-exemples actionnables (correspondant à l'étape N° 3 de notre méthodologie) et profitables (correspondant à l'étape N° 4 de notre méthodologie) est réalisée exclusivement dans la classe *Cohorte*.

5.1.3.2 Cas d'utilisation

Le diagramme de cas d'utilisation présenté en figure 5.7 permet de distinguer le rôle des deux acteurs principaux de l'application *ARKIS* : le *data miner* et l'expert métier. Le *data miner* est chargé des étapes de pré-traitement jusqu'à la génération des cohortes. L'expert métier filtre et valide les règles et cohortes puis paramètre l'outil en fonction des besoins du métier, pour extraire les contre-exemples actionnables et profitables et ainsi déclencher les recommandations.

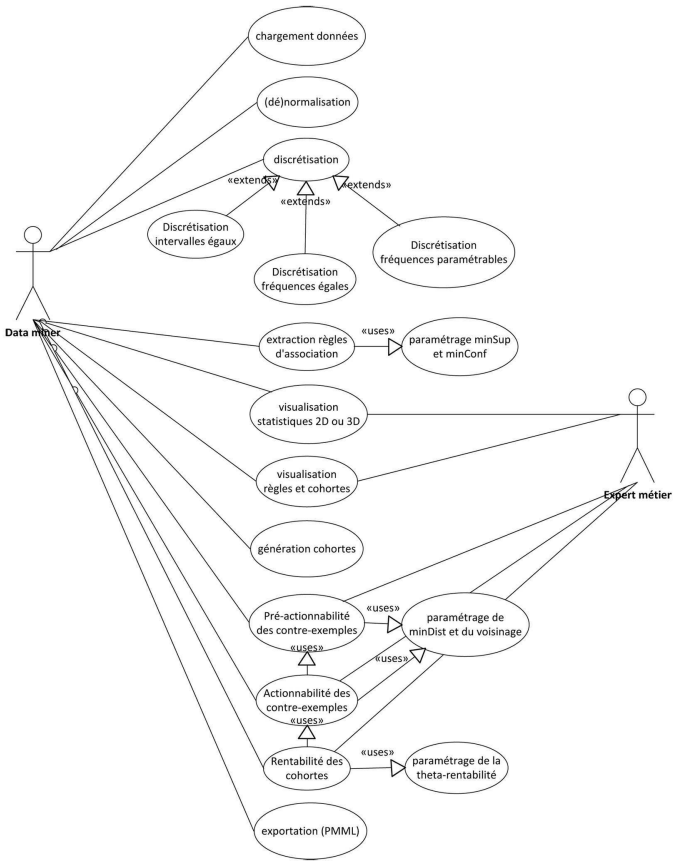


FIGURE 5.7 : Diagramme de cas d'utilisation d'*ARKIS*

5.1.4 Exemple d'utilisation

Faisant référence au schéma d'architecture générale d'*ARKIS* (cf. figure 5.1), nous présentons dans l'exemple suivant la génération de la matrice des distances à partir des variables descriptives et l'extraction des recommandations actionnables et profitables à partir des variables d'achats.

L'exemple d'utilisation présenté ci-dessous a été réalisé avec la version 1.0 de l'application *ARKIS*. Les données étudiées correspondent à celles illustrant notre méthodologie dans le chapitre 4. Rappelons que la base de données est composée d'achats de dix clients $\{c_1, c_2, \ldots c_{10}\}$ pour cinq produits $\{p_1, p_2, \ldots p_5\}$. De plus, chaque client c est décrit par son âge, son sexe et son revenu annuel.

Génération de la matrice des distances

La figure 5.8 reprend la partie du schéma 5.1 présentant la génération de la matrice des distances :

FIGURE 5.8 : Processus de génération de la matrice des distances

D'un point de vue applicatif, il suffit de charger le fichier comprenant les variables descriptives des clients (cf. figure 5.9). La matrice des distances triangulaire est générée et stockée dans un fichier CSV (cf. figure 5.10).

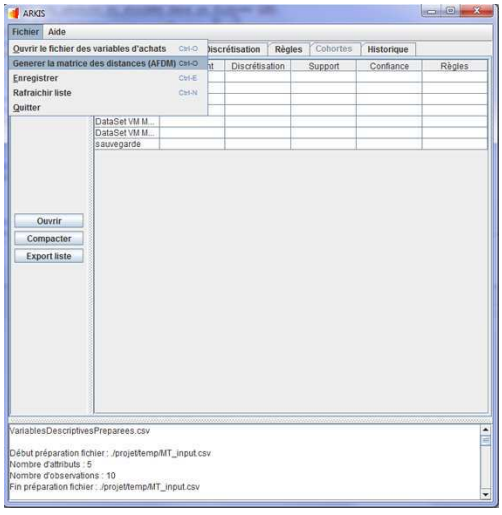


FIGURE 5.9 : Chargement du fichier des variables descriptives des clients

DISTANCES	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
c1	0,00000									
c2	0,15061	0,00000								
c3	8,33320	6,29994	0,00000							
c4	7,11876	5,19846	0,43785	0,00000						
c5	8,75667	6,61045	0,43785	0,08472	0,00000					
c6	4,97268	3,46656	0,46125	0,70142	1,09677	0,00000				
c7	3,39818	2,11798	1,35799	0,68011	1,24490	0,50139	0,00000			
c8	4,76794	3,29947	0,52949	0,75554	1,17914	0,00235	0,47550	0,00000		
c9	5,83869	4,18196	0,23533	0,53198	0,81438	0,03765	0,65200	0,05883	0,00000	
c10	6,12096	4,35127	0,53198	0,03765	0,23533	0,53198	0,39771	0,56728	0,43785	0,0000

FIGURE 5.10 : Visualisation de la matrice des distances entre clients

Extraction des recommandations

Le processus d'extraction de recommandations actionnables et profitables est présenté sur la figure 5.1. Plusieurs étapes de pré-traitement sont réalisées : chargement du fichier, traitement des valeurs nulles et négatives, normalisation et discrétisation (cf. figures 5.11).

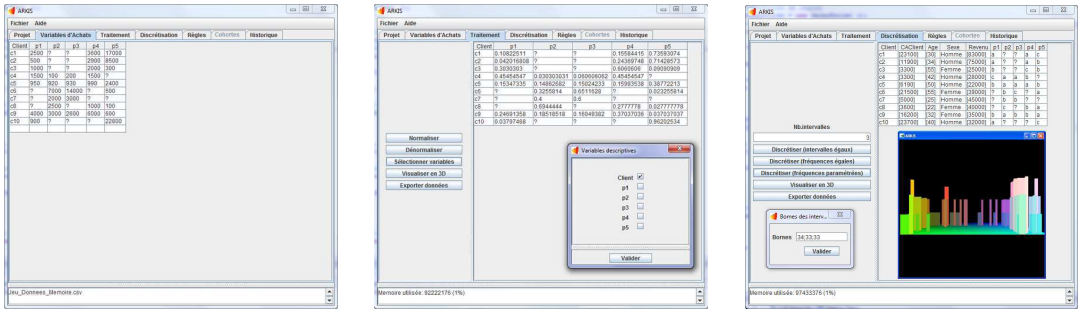


FIGURE 5.11 : Traitement, normalisation et discrétisation des variables d'achats

L'ensemble des règles d'association est ensuite extrait en fonction des paramètres $minSup$ et $minConf$ (cf. figure 5.12). Quelques statistiques des variables descriptives sont présentées. Les cohortes peuvent être ensuite générées (cf. figure 5.13).

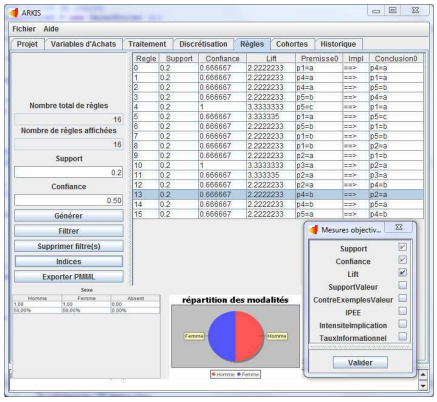


FIGURE 5.12 : Extraction des règles d'association

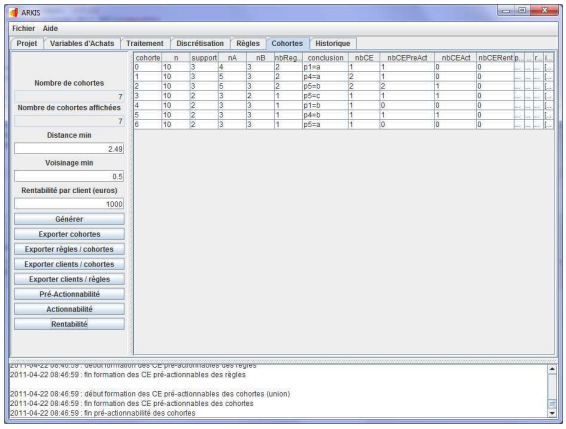


FIGURE 5.13 : Génération des cohortes, pré-actionnabilité, actionnabilité et profitabilité

L'étape de pré-actionnabilité cible, pour chaque cohorte, les contre-exemples pré-actionnables en fonction des variables d'achats sur les éléments de la prémisse de la cohorte. L'étape d'actionnabilité permet de cibler davantage les clients en utilisant la matrice des distances générée dans le processus précédent. Par exemple, sur la figure 5.13, l'utilisateur souhaite que chaque contre-exemple pré-actionnable possède 50 % des exemples de la cohorte à une distance inférieure ou égale à 2,49 (moyenne sur la matrice des distances). Enfin, l'utilisateur peut extraire les contre-exemples profitables en fonction d'un seuil θ égal à 1 000 Euros à partir duquel un contre exemple actionnable est considéré comme profitable.

5.2 Expérimentation et validation sur les données *MovieLens*

5.2.1 Données *MovieLens*

Les votes

Afin d'évaluer la qualité de notre méthodologie de recommandations actionnables et profitables, nous utilisons le jeu de données *MovieLens* [18] fournie par l'équipe de recherche américaine *GroupLens* [19]. *MovieLens* [20] est un site Internet de recommandations de films. Les utilisateurs ont la possibilité de partager leurs préférences en votant explicitement pour des items sur une échelle de valeurs entières comprises entre 1 et 5. Le jeu de données *MovieLens* a été largement utilisé par la communauté scientifique pour évaluer et comparer les algorithmes de filtrage collaboratif [193]. Il présente en effet l'avantage de reposer sur des votes réels et fournit de ce fait un bon support de validation. La matrice contient 943 utilisateurs (C), 1 682 items (P) et 100 000 votes (U). Ainsi, la matrice de votes présente une dispersion de 6,30 % (cf. section 3.2.6.1), c'est-à-dire qu'il y a 93,70 % de données manquantes, considérées comme des non-votes. La distribution des votes est présentée dans le tableau 5.5. Une grande proportion des votes sont des 3 et 4 (votes médians ou satisfaisants). Il y a globalement peu de votes d'utilisateurs insatisfaits (1 ou 2). Le jeu de données a été divisé en cinq ensembles d'apprentissage et cinq ensembles de test appelés respectivement « base » et « test ».

Votes / Jeux de données	1	2	3	4	5
U.data	6,11 %	11,37 %	27,14 %	34,17 %	21,20 %
U1.base	5,90 %	11,47 %	27,45 %	34,25 %	20,93 %
U2.base	6,06 %	11,48 %	27,26 %	34,12 %	21,07 %
U3.base	6,15 %	11,43 %	26,95 %	34,05 %	21,40 %
U4.base	6,24 %	11,21 %	26,97 %	34,26 %	21,31 %
U5.base	6,19 %	11,24 %	27,08 %	34,19 %	21,29 %
U1.test	6,96 %	10,96 %	25,91 %	33,90 %	22,27 %
U2.test	6,29 %	10,93 %	26,67 %	34,41 %	21,71 %
U3.test	5,94 %	11,10 %	27,90 %	34,66 %	20,40 %
U4.test	5,59 %	11,98 %	27,85 %	33,83 %	20,75 %
U5.test	5,79 %	11,89 %	27,39 %	34,10 %	20,83 %

TABLEAU 5.5 : Distribution des votes sur les données *MovieLens*

U.data correspond au jeu de données complet. *U[1-5].base* sont les cinq ensembles d'apprentissage et *U[1-5].test* sont les cinq ensembles de validation générés. Les ensembles d'apprentissage et de test contiennent respectivement 80 % et 20 % des votes globaux.

18. http://www.grouplens.org/node/73
19. http://www.grouplens.org
20. http://www.movielens.org/login

Les variables descriptives

Les 943 clients du jeu de données sont décrits par trois variables descriptives qualitatives (*sexe*, *métier*, *code postal*) et une variable quantitative (*âge*) :

- Les clients sont représentés par 670 hommes et 273 femmes (cf. figure 5.14) ;
- Les métiers représentatifs sont les étudiants (196, *student*), les autres (105, *other*), les professeurs (95, *educator*) et les agents de l'administration (79, *administrator*) (cf. figure 5.15) ;
- La répartition des codes postaux est davantage dispersée avec 9 individus dans le 55414, 6 individus dans le 55105, 5 individus pour les codes 10003, 20009 et 55337 (cf. figure 5.16). Cette variable apportant peu d'information n'est pas sélectionnée comme variable descriptive des clients ;
- La répartition des âges se distribue entre 7 et 73 ans, avec une médiane de 31 ans et une moyenne de 34 ans. Les quartiles $Q1$ et $Q3$ soulignent la concentration de 50 % des individus entre 25 et 43 ans. Enfin, la variance est de 148,66 et l'écart-type de 12,19 (cf. figure 5.17).

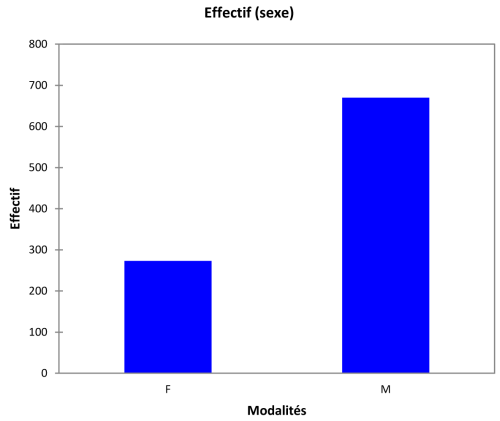


FIGURE 5.14 : Diagramme en bâtons de la variable *sexe*

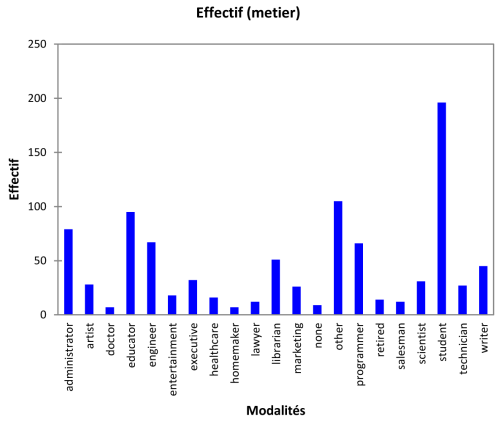


FIGURE 5.15 : Diagramme en bâtons de la variable *metier*

5.2.2 Choix effectués

Le principe est d'exploiter l'ensemble d'apprentissage pour générer des prédictions de recommandations à partir de notre méthodologie *CAPRE* et de comparer les résultats avec les valeurs réelles contenues dans l'ensemble de validation. L'opération est répétée cinq fois afin de renforcer la validité de l'évaluation statistique. Dans la mesure où les votes sont séparés aléatoirement dans les ensembles « base » et « test », la qualité des prédictions peut varier en fonction de la proportion de votes représentatifs contenus dans l'ensemble d'apprentissage.

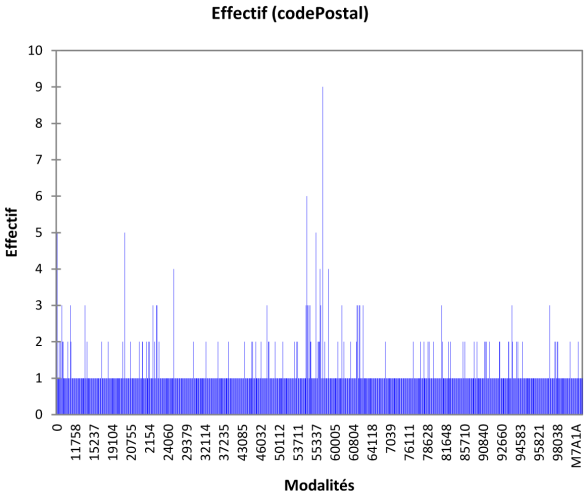


FIGURE 5.16 : Diagramme en bâtons de la variable *codePostal*

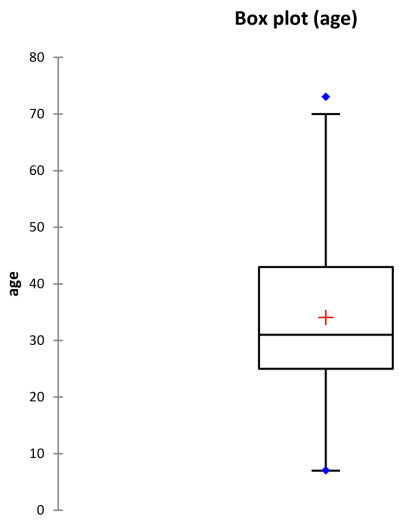


FIGURE 5.17 : *Box plot* de la variable *âge*

Dans le cadre de notre méthodologie et notamment à l'étape de profitabilité, nous introduisons une notion de chiffre d'affaires dans le jeu de données *MovieLens*. Pour chaque client actionnable de la méthodologie, nous considérons que la recommandation rapporte 5 €[21]. Chaque utilisateur a voté pour au moins 20 films sur une échelle de 1 (insatisfait) à 5 (satisfait). Le jeu de données a été transformé en une matrice $C \times P$ composée de 943 lignes (les clients) et de 1 682 colonnes (les films votés par au moins un client). Les variables descriptives des clients, c'est-à-dire $N \cup I$ sont l'âge, le sexe et le métier. Les votes sont regroupés en trois classes $a \in [1,2]$, $b \in [3]$ et $c \in [4,5]$ permettant ainsi d'identifier respectivement les clients insatisfaits, mitigés et satisfaits au regard du film voté. Nous considérons que les cases vides de la matrice représentent des clients n'ayant pas visualisé le film en question.
Notre objectif est de recommander aux clients actionnables et profitables (c'est-à-dire susceptibles d'apprécier la recommandation) la visualisation d'un ou plusieurs films à l'aide de notre méthodologie *CAPRE*.

5.2.3 Exemple d'application de la méthodologie *CAPRE*

Avec un seuil de support $minSup$ de 4 % (environ 40 clients) et un seuil de confiance $minConf$ égal à 50 %, **26 658** règles d'association sont extraites (cf. tableau 5.6) sur le jeu de données *U1.Base*.
Considérons la règle *Men in Black* = $c \rightarrow$ *Indep._Day* = c. 102 clients qui ont vu et voté positivement (4 ou 5) pour le film d'action *Men_in_Black* ont également vu et appréciés le film d'action *Indep._Day*. Cependant, 43 contre-exemples ont visualisé et voté positivement pour le film *Men_in_Black* mais n'ont pas vu le film *Indep._Day*.

21. Le prix moyen actuel de la visualisation d'un film en *streaming* sur Internet est de 5 €. Par exemple, le tarif d'entrée sur le site de *NetFlix* est de $7.99.

Support (%)	**Confiance (%)**	**Règles**
0.11 %	0.54 %	Men_in_Black = c → Indep._Day = c
0.12 %	0.67 %	Die_Hard = c → The_Terminator = c
0.10 %	0.76 %	Mission_Impossible = b → Star_Wars = c
0.12 %	0.83 %	The_Shining = c → Silence_of_the_Lambs = c
0.12 %	0.79 %	Speed = b → The_Fugitive = c
0.10 %	0.73 %	Beauty_and_the_Beast = c → Toy_Story = c

TABLEAU 5.6 : Exemples de règles d'association extraites sur *U1.base*

L'application de la suite de la méthodologie *CAPRE* permet de générer un ensemble de **99** cohortes de règles d'association, c'est-à-dire 99 recommandations de films. Concentrons notre analyse sur la cohorte suggérant le film *Indep._Day* (cf. tableau 5.7). La cohorte est composée de quatre règles d'association.

Support (%)	**Confiance (%)**	**Règles**
0.13 %	0.54 %	The_Rock = c → Indep._Day = c
0.12 %	0.52 %	Star_Trek = c → Indep._Day = c
0.11 %	0.54 %	Men_in_Black = c → Indep._Day = c
0.10 %	0.56 %	Star_Wars = c and The_Rock → Indep._Day = c

TABLEAU 5.7 : Quatre règles d'association de la cohorte $Ct(Indep.\_Day = c)$

La cohorte *Ct(Indep._Day=c)* est composée de 192 clients exemples et 188 clients contre-exemples. Les étapes de pré-actionnabilité, d'actionnabilité et de profitabilité permettent d'élaguer le nombre de recommandations et de clients ciblés de la manière suivante :

- *Pré-actionnabilité* : de 188 à 140 clients contre-exemples pré-actionnables. Par exemple, le client #236 n'est pas pré-actionnable car il apprécie le film *Star_Trek = c* mais est insatisfait par les films *The_Rock*, *Men_in_Black* et *Star_Wars* ;

- *Actionnabilité* : de 140 à 93 clients contre-exemples actionnables. Nous remarquons que les exemples sont principalement des hommes entre 20 et 45 ans, occupant une profession d'étudiant (*student*), de responsable (*executive*) ou d'ingénieur (*engineer*). Dès lors, nous élaguons de nombreux contre-exemples, tels que les femmes écrivains ou artistes de 25 ans ;

- *Profitabilité* : en assignant une valeur économique de 5 € à chaque film et en considérant la recommandation du film *Independence_Day*, nous estimons une profitabilité de 465 € pour la recommandation de la cohorte. Dès lors, ce profit espéré peut être comparé aux profits de l'ensemble des 99 cohortes générées.

5.2.4 Discussion

La méthodologie *CAPRE* permet (i) d'élaguer largement le nombre de recommandations à actionner par les experts métier et (ii) de filtrer pertinemment le nombre de clients potentiellement actionnables et profitables. De plus, nous observons des comportements de votes clients à travers d'autres cohortes :

- Certaines cohortes présentent un ensemble d'exemples dispersé, dû en partie, à la présence de films très populaires tel que *Star_Wars* pouvant toucher de nombreuses typologies de clients. Le comportement des exemples est difficile à caractériser à travers les votes ou les variables descriptives. C'est pourquoi, nous élaguons peu les contre-exemples. L'ensemble des contre-exemples actionnables est proche de l'ensemble des contre-exemples de la cohorte.
- D'autres cohortes, relatives aux films d'horreurs, présentent des exemples possédant des variables descriptives semblables (cf. figure 5.18). Les contre-exemples qui ont vu et apprécié au moins un film d'horreur seront généralement facilement identifiables pour les étapes d'actionnabilité et de profitabilité. Dès lors, la frontière entre le nuage d'exemples et le nuage de contre-exemples actionnables se dessine plus aisément pour ce type de cohorte.

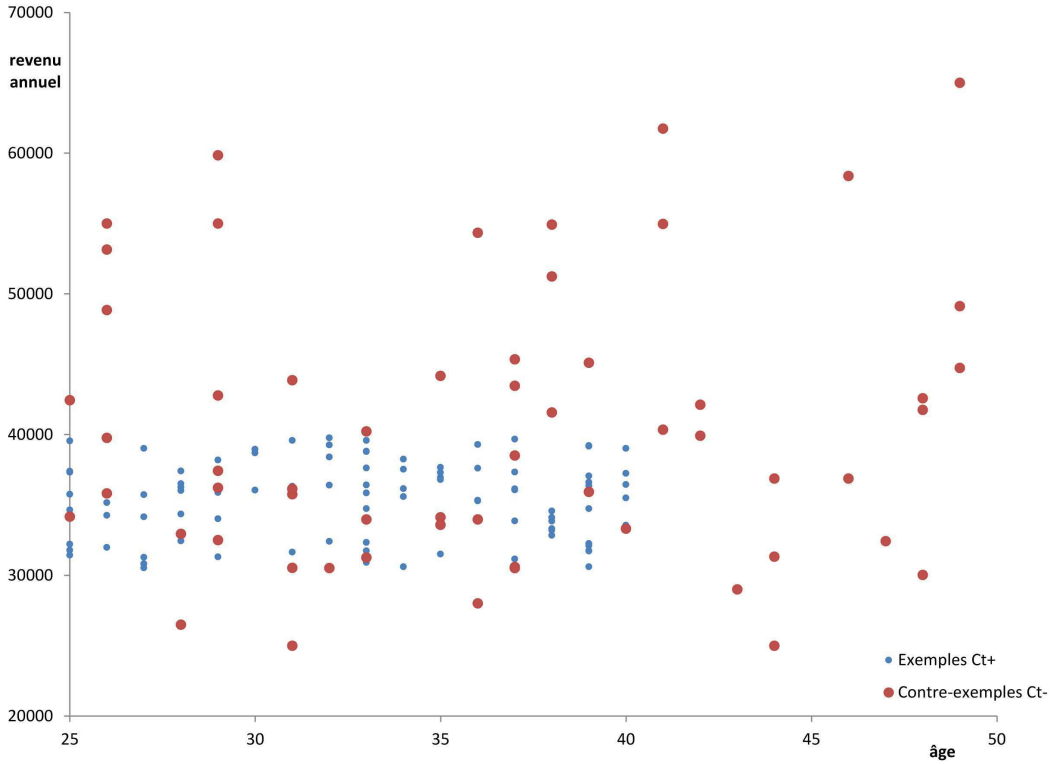


FIGURE 5.18 : Nuage de points des exemples concentrés dans un espace 2D des variables descriptives (âge et revenu annuel)

Enfin, Onuma et al. [208] recommandent pour les adeptes des films d'horreur (*Horror Movie Fan*) des films tels que : *A_Nightmare_on_Elm_Street*, *The_Shining* and *Jaws*. Cependant, avec notre méthodologie *CAPRE*, nous remarquons que pour la cohorte recommandant le film *Jaws*, composée de sept règles d'association, nous élaguons les 144 contre-exemples de la cohorte pour obtenir 75 contre-exemples actionnables. Par conséquent, nous ciblons plus finement la recommandation aux clients estimés les plus réceptifs à regarder le film *Jaws*. Finalement, si nous considérons un gain estimé à 5 € par visualisation du film et un gain publicitaire sur le site Internet de *MovieLens*, un retour sur investissement de la recommandation peut être estimé.

5.2.5 Validation croisée et comparaison

Tout d'abord, intéressons-nous à la validation des résultats sur les cinq ensembles d'apprentissage et les cinq ensembles de validation (cf. section 5.2.1). Nous réalisons une extraction des règles avec un support $minSup$ égale à 0,04 (environ 38 clients) et une confiance $minConf$ égale à 50 %. Nous générons par la suite un ensemble de cohortes. Les premiers résultats sont récapitulés dans le tableau 5.8.

	U1.base	**U2.base**	**U3.base**	**U4.base**	**U5.base**
# de règles	26 658	12 487	12 456	9 516	13 028
# de cohortes	99	95	92	88	92

TABLEAU 5.8 : Application de *CAPRE* sur les données *MovieLens*

Pour maintenir une certaine robustesse, les cohortes composées d'au moins 20 règles sont conservées (cf. figure 5.19). Afin d'évaluer les *Top-N* recommandations , nous utilisons deux métriques classiquement utilisées dans la communauté des systèmes de recommandation, à savoir, la précision et le rappel [75]. Nous privilégions la mesure de précision afin de minimiser le nombre de faux positifs (cf. tableau 5.9), c'est-à-dire les clients que nous détecterions actionnables mais qui dans la réalité n'ont pas apprécié le film. [75]. Enfin, nous présentons dans le tableau 5.11 les *Top-10* et les *Bottom-10* recommandations triées par précision moyenne $\overline{Pr}$ décroissante.

Conformément à nos objectifs centrés sur la fidélité et la satisfaction client, nous avons construit la matrice 5.9 pour juger de l'impact des recommandations sur les critères de fidélisation et de satisfaction.

		Réel	
		Votes 4 ou 5	Votes 1, 2 ou 3
Prédit	Actionnable $c \in [4,5]$ Non Actionnable $a \in [1,2]$ ou $b \in [3]$	↗ fidélisation	↘ satisfaction
		↘ fidélisation	↗ satisfaction

TABLEAU 5.9 : Matrice des impacts de la recommandation sur la relation client

Nous remarquons que le fait de proposer un film à un client qui ne va pas l'apprécier peut impacter sa satisfaction envers le système. De la même manière, ne pas proposer à un client un film susceptible de l'intéresser peut impacter sa fidélité vis-à-vis du système. Ce point de vue offre la possibilité aux experts métier d'élaborer leur politique de recommandation. Nous obtenons de bons résultats à l'aide de la mé-

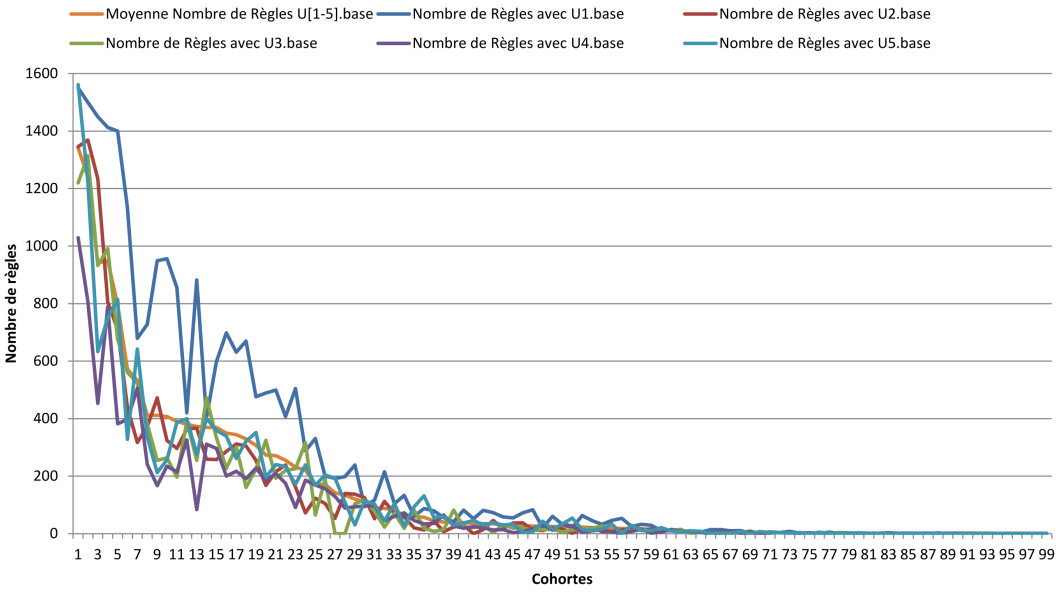


FIGURE 5.19 : Évolution du nombre de règles par cohorte sur *U[1-5].base*

thodologie *CAPRE* sur la précision moyenne des *Top-10* recommandations (même les résultats sur les *Bottom-10* sont acceptables). Les erreurs mesurées par la précision et le rappel sont similaires à ceux généralement mesurées sur le jeu de données *MovieLens* avec par exemple la mesure RMSE (*Root Mean Square Error*). L'agrégation des règles en cohortes permet de construire un modèle plus robuste qu'une simple extraction de règles d'association, réduisant ainsi les erreurs de recommandations.

5.2.6 Impact de l'actionnabilité sur la précision

L'impact de l'actionnabilité sur les *Top-10* et les *Bottom-10* recommandations (cf. tableau 5.10) souligne que les phases de pré-actionnabilité et d'actionnabilité permettent de diminuer fortement le nombre de contre-exemples à démarcher. En effet, le nombre de contre-exemples actionnables est diminué en moyenne de 19,08 %. Néanmoins, il est remarquable de constater que la précision moyenne se maintient ou augmente ($\simeq +1,03$ %). Dès lors, le nombre de faux positifs est élagué tandis que le nombre de vrais positifs est maintenu. Par exemple, pour la recommandation du film *La Liste de Schindler* (*Schindlers List*), 34 clients contre-exemples sont élagués. La précision moyenne est améliorée de 0,70 %. De la même manière, le nombre d'utilisateurs réceptifs à la recommandation du film Terminator (*The Terminator*) est élagué de 37 individus tandis que la précision est augmentée de 0,41 % et atteint les 72,87 %.

Top-10	Avant l'actionnabilité		Après l'actionnabilité	
	Nombre de contre-exemples	Précision moyenne	Nombre de contre-exemples actionnables	Précision moyenne
Rear Window	199	90,79	155	91,50
The Shawshank Redemption	241	90,05	201	91,07
Casablanca	245	89,74	189	90,02
Schindlers List	235	89,27	201	89,97
To Kill a Mockingbird	194	89,00	167	89,03
One Flew Over Cuckoos Nest	242	87,22	207	88,28
The Silence of the Lambs	276	87,35	227	87,98
The Usual Suspects	257	87,07	210	86,76
Star Wars	262	86,08	221	86,58
The Godfather	258	84,09	224	85,01
Bottom-10				
The Terminator	247	72,46	210	72,87
2001 : A Space Odyssey	185	72,30	157	72,32
Toy Story	275	70,98	235	71,48
Die Hard	154	68,59	112	71,36
Contact	254	69,45	201	70,85
Indiana Jones the Last Crusade	253	68,89	219	69,81
Forrest Gump	257	69,14	220	69,73
Twelve Monkeys	244	67,51	179	68,04
Back to the Future	257	67,32	213	67,83
Jaws	144	66,63	75	67,29

TABLEAU 5.10 : Impact de l'actionnabilité sur la précision moyenne des *Top-10* et *Bottom-10* recommandations

5.3 Conclusion

Dans ce chapitre, nous avons présenté l'outil de recommandations actionnables et profitables *ARKIS*, une implémentation de la méthodologie *CAPRE*. *ARKIS* est un outil abouti et structuré permettant de répondre aux sept étapes de la méthodologie *CAPRE* à partir des variables d'achats et des variables descriptives des clients.
La validation de la méthodologie *CAPRE* à l'aide de l'outil *ARKIS* sur les données de référence *MovieLens* montre des recommandations intéressantes et de bonne qualité. En particulier, les cohortes avantagent la robustesse du système, réduisant ainsi les fausses recommandations.
Dans la prochain et dernier chapitre de la thèse, nous présentons une application de la méthodologie *CAPRE* sur la base de données réelle de VM Matériaux composée de plus de 10 000 clients et 100 000 produits.

Top-10	U1.test		U2.test		U3.test		U4.test		U5.test		**Moyenne**	
	Pr	Rp	Pr	Rp	Pr	Rp	Pr	Rp	Pr	Rp	$\overline{\mathbf{Pr}}$	$\overline{\mathbf{Rp}}$
Rear Window	96,00	70,59	88,24	83,33	87,10	72,97	96,97	86,49	89,19	80,49	**91,50**	78,77
The Shawshank Redemption	93,75	78,95	86,00	87,76	90,91	86,96	93,62	93,62	91,07	91,07	**91,07**	87,67
Casablanca	97,37	82,22	83,67	85,42	92,31	87,80	88,10	92,50	88,64	92,86	**90,02**	88,16
Schindlers List	86,54	86,54	89,29	90,91	91,84	93,75	90,32	91,80	91,84	91,84	**89,97**	90,97
To Kill a Mockingbird	83,33	75,76	96,55	71,79	90,91	78,95	84,62	84,62	89,74	79,55	**89,03**	78,13
One Flew Over Cuckoos Nest	91,49	91,49	87,23	83,67	85,71	94,12	89,19	89,19	87,76	93,48	**88,28**	90,39
The Silence of the Lambs	93,24	84,15	89,71	88,41	81,94	89,39	88,24	96,77	86,76	90,77	**87,98**	89,90
The Usual Suspects	85,71	87,50	86,27	81,48	82,22	88,10	89,13	93,18	90,48	86,36	**86,76**	87,32
Star Wars	87,95	84,88	84,68	90,38	82,79	90,18	89,72	90,57	87,76	92,47	**86,58**	89,70
The Godfather	85,25	82,54	84,29	86,76	85,90	94,37	82,09	84,62	87,50	83,33	**85,01**	86,32
Bottom-10												
The Terminator	64,29	85,71	77,78	83,33	73,77	95,74	66,67	92,31	81,82	95,74	**72,87**	90,57
2001 : A Space Odyssey	81,40	76,09	71,43	85,71	68,57	96,00	71,43	83,33	68,75	84,62	**72,32**	85,15
Toy Story	70,37	77,55	70,93	95,31	74,16	90,41	67,50	87,10	74,42	87,67	**71,48**	87,61
Die Hard	81,25	72,22	61,76	80,77	85,29	67,44	72,09	93,94	56,41	75,86	**71,36**	78,05
Contact	66,67	61,97	85,11	52,63	70,69	58,57	65,12	83,58	66,67	76,67	**70,85**	66,68
Indiana Jones the Last Crusade	67,92	85,71	64,71	88,00	68,63	85,37	67,80	93,02	80,00	94,92	**69,81**	89,40
Forrest Gump	67,31	83,33	77,05	92,16	72,88	93,48	71,43	93,02	60,00	83,72	**69,73**	89,14
Twelve Monkeys	76,19	78,69	61,54	86,96	66,67	80,00	69,12	90,38	66,67	85,71	**68,04**	84,35
Back to the Future	70,77	86,79	68,85	93,33	60,32	92,68	66,18	95,74	73,02	93,88	**67,83**	92,48
Jaws	70,83	80,95	78,38	70,73	52,73	93,55	69,39	94,44	65,12	82,35	**67,29**	84,40

TABLEAU 5.11 : Précicion (Pr) et rappel (Rp) des *Top-10* et *Bottom-10* recommandations sur le jeu de données *MovieLens*

6

Validation sur les données VM Matériaux

It is important to develop economics-oriented measures that capture the business value of recommendations [...]

Gediminas Adomavicius, 2005

SOMMAIRE

6.1 Phase préparatoire : ciblage des opérations commerciales

6.1.1 Contexte

Les activités du Négoce de matériaux représentent un marché extrêmement concurrentiel. Pour les acteurs de ce marché, l'optimisation de la gestion de la relation client à travers des méthodes de fouille de données peut s'avérer être un élément différenciateur, permettant ainsi de dégager une satisfaction et une fidélisation client, et des gains de rentabilité importants.
L'activité Négoce du groupe VM Matériaux organise en septembre une semaine commerciale promotionnelle réservée aux professionnels du bâtiment et destinée à promouvoir l'ensemble des produits. Les clients peuvent ainsi gagner des points VM Matériaux ou différents cadeaux, ceci sous réserve de la réalisation d'un CA passé en commande sur cette période et facturé en fin de mois. Habituellement, seuls les clients professionnels ayant réalisé un minimum de CA l'année précédente sont démarchés par un courrier et par leurs commerciaux respectifs. Un commercial gère un portefeuille de clients professionnels sur lequel des objectifs lui sont affectés. Dès lors, le commercial représente le point d'entrée essentiel pour la relation avec les clients professionnels.
Dans ce chapitre, nous présentons le retour d'expérience du projet de fouille de données mené chez VM Matériaux pour améliorer le retour sur investissement d'opérations commerciales. Notre objectif est de réaliser un ciblage des clients susceptibles de participer à l'opération commerciale (cf. de la section 6.1.2 à la section 6.1.10). Ce ciblage permet de sélectionner la population de clients professionnels la plus appétente pour l'application des recommandations de notre méthodologie *CAPRE* (cf. section 6.2). La réalisation de la phase préparatoire, c'est-à-dire le ciblage de l'opération commerciale, se décompose en quatre étapes instanciées à l'aide de l'outil de fouille de données *KXEN* (cf. annexe C) :

Étape N° 1
Ciblage de l'appétence client pour l'opération commerciale (*cible binaire*)

Étape N° 2
Estimation de la marge nette par client participant (*cible continue*)

Étape N° 3
Choix des canaux de communication adapté au profil estimé par client

Étape N° 4
Mise en place d'un modèle économique de retour sur investissement

6.1.2 Interactivité avec les experts métier

L'implication des experts du domaine et l'apport de leurs connaissances peuvent réduire la complexité du processus d'extraction de connaissances actionnables. Cette sous section met en lumière l'importance des décideurs métier tout au long du processus. Nous avons formé une équipe de travail composée d'experts du métier et des données : le directeur de l'activité Négoce (Jean-Charles Chaîgne), le directeur marketing Négoce (Philippe Queneau), le chef de projet Négoce métier (Yann Froment), le directeur projets groupe (Patrice Vequeau), le directeur informatique (Pierrick Richard), l'adjoint au directeur informatique (Gaëtan Blain) et l'équipe force de vente de VM Matériaux, contribuant ainsi à la transformation de la donnée en information et de l'information en connaissance. L'ensemble des étapes correspondantes au concept « *Human Cooperated Mining* » définies dans la méthodologie *DDID-PD* présentée dans la section 2.2.2.2 ont été respectées par la cellule de travail, à savoir :

- la définition de la problématique métier ;
- la compréhension des données ;
- l'intégration et l'échantillonnage des données ;
- la modélisation et la compréhension du modèle ;
- l'interprétation et l'amélioration du modèle pour l'interprétation des variables ;
- la comparaison des résultats ;
- le choix des canaux de communication marketing ;
- le calcul du retour sur investissement.

Ces étapes ont été simplifiées sur la figure 6.1 et la répartition du temps passé par les experts est indiquée. Chaque étape donne souvent lieu à de nombreuses itérations avec les étapes précédentes.

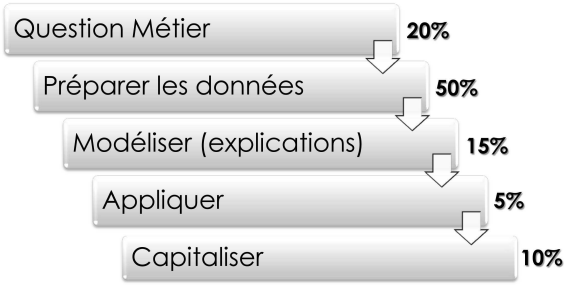


FIGURE 6.1 : Étapes clefs et répartition du temps passé avec les experts métier

6.1.3 Pré-traitement

Préparation des données

Nous avons extrait les données à partir de l'entrepôt existant composé approximativement de 180 tables pour un volume de 93 gigaoctets (cf. vue simplifiée de l'entrepôt de données en annexe A). Considérons un tableau (cf. tableau 6.1) de clients pour lesquels la variable cible binaire illustre l'achat durant la campagne précédente et la variable cible continue correspond à la marge nette générée par la visite du client durant l'opération commerciale.

Identifiant client	Cible binaire (achat / non achat)	Cible continue (marge nette en €)
PR00001	1	555,20
PR00002	0	NULL
PR00003	1	269,25
...	...	...

TABLEAU 6.1 : Définitions des cibles binaire et continue des clients

Trois jeux de données distincts ont été créés : le jeu de données d'apprentissage pour une cible binaire, le jeu de données d'apprentissage pour une cible continue et le jeu de données d'application.

- Le jeu de données d'apprentissage pour une cible binaire est l'ensemble de tous les clients actifs professionnels de la table client de l'entrepôt de données. La cible binaire prend la valeur 1 lorsque les clients ont participé à la dernière campagne marketing pour un seuil de chiffre d'affaires de 1 500 € (déterminé par les experts métier), et 0 dans les autres cas.

- Le jeu de données d'apprentissage pour une cible continue est composé de l'ensemble de tous les clients professionnels ayant participé à la dernière campagne marketing (au moins un achat), les autres clients étant supprimés. La cible continue représente la marge nette [1] en Euros réalisée par les clients lors de la dernière campagne. La taille de ce jeu de données est inférieure au jeu de données précédent.

- Le jeu de données d'application repose sur les clients professionnels filtrés par le directeur commercial et l'équipe forces de vente. Plus précisément, sont sélectionnés les clients mouvementés (réalisant un seuil de chiffre d'affaires sur 12 mois glissants), qui ne sont pas des clients grands comptes (conseils municipaux, organismes gouvernementaux, etc.) et qui ne présentent pas de problème majeur de crédit.

Les caractéristiques des trois jeux de données présentés précédemment sont résumées dans le tableau 6.2.

1. La marge nette correspond à la différence entre le chiffre d'affaires net et le prix de revient. Elle constitue un élément essentiel pour le ciblage de l'opération commerciale.

Jeu de données	Nombre d'observations	Caractéristiques
Apprentissage cible binaire	12 170	15.35 % de valeur cible à 1
Apprentissage cible continue	3 677	5 400 € de CA moyen
Application	19 000	Clients mouvementés

TABLEAU 6.2 : Caractéristiques des trois jeux de données

Création et sélection des variables

Les exigences et les problématiques des entreprises sont souvent étroitement liées aux règles spécifiques du domaine (cf. concept « *Including Constraint Mining* » défini dans la méthodologie *DDID-PD* présentée dans la section 2.2.2.2). Dès lors, les clients professionnels sont décrits par trois types de variables qui seront interprétées en fonction des connaissances du métier :

- **Variables internes** (dans l'entrepôt de données) : adresse, métier, fidélité, type de profil (mouvementé, endormi, prospect), adhérence (carte de fidélité), ancienneté, commercial principal, magasin de rattachement, encours autorisé, etc.
- **Variables externes** (généralement obtenues à l'aide d'un fichier *Coface* [2]) : nombre d'employés, catégorie socio-professionnelle, bilan, etc.
- **Agrégats** (calculés dans le SGBD pour chaque client) :
  - Somme de marge nette sur la dernière campagne similaire ;
  - Somme de chiffre d'affaires et de lignes de commandes sur 12 périodes de un mois avant la campagne ;
  - Somme de chiffre d'affaires pour chaque famille de produits (gros-œuvre, couverture, étanchéité, etc.) sur 12 mois glissants avant la campagne.

Cette étape de pré-traitement des données permet de construire un modèle de données composé de 266 variables (cf. tableau 6.3).

Catégorie	Nombre	Type
Variables internes	20	Numériques (*encours*) et catégoriques (*métier*)
Variables externes	19	Données non-consolidées (*bilan de l'entreprise*)
Agrégats	227	Continus (*sommes de chiffre d'affaires ou marge nette*)

TABLEAU 6.3 : Types de variables du modèle de données

2. http ://www.coface.com

6.1.4 Génération des modèles

Le score de tendance utilisant la régression *ridge* permet d'estimer la probabilité de chaque client à participer durant la campagne marketing. Les experts métier valident les modèles à l'aide de courbes *lift* et de courbes de profit construites respectivement à l'aide d'une matrice de confusion et d'une matrice des coûts. L'évaluation des modèles à l'aide des indices KI et KR est expliquée en annexe C.
Le modèle pour la cible binaire est précis et robuste avec un $KI = 0,889$ et un $KR = 0,975$. 260 des 266 variables sont sélectionnées, préservant ainsi la précision et la robustesse du modèle. Les courbes *lift* et ROC (cf. figure 6.2) permettent de juger de la qualité du modèle.

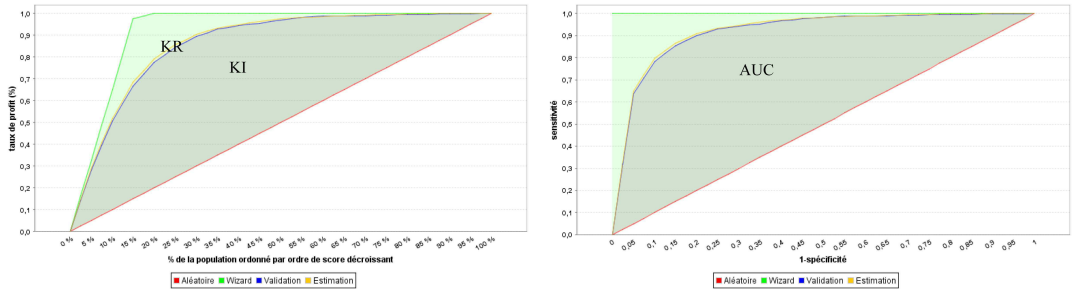


FIGURE 6.2 : Courbes *lift* (à gauche) et ROC (à droite) du modèle à cible binaire

Le modèle pour la cible continue est moins précis et moins robuste avec un $KI = 0,717$ et un $KR = 0,968$, notamment pour les « gros » clients présentant un chiffre d'affaires significatif (cf. figure 6.3).

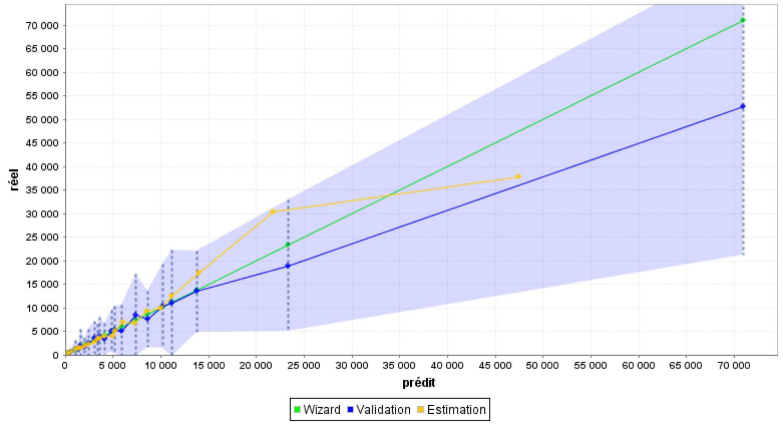


FIGURE 6.3 : Prédit / réel pour le modèle à cible continue

Les experts métier privilégient le ciblage d'un nouveau client (Étape N° 1) par rapport au développement de la marge nette (Étape N° 2). Par conséquent, les listes de ciblage communiquées aux commerciaux sont triées prioritairement sur la probabilité de participation. Dans la suite de l'application, nous nous concentrons davantage sur l'explication du modèle pour la cible binaire.

6.1.5 Interprétation

Visualisons la contribution des variables pour distinguer les variables les plus contributrices à l'achat (cible binaire) durant la campagne marketing (cf. figure 6.4).

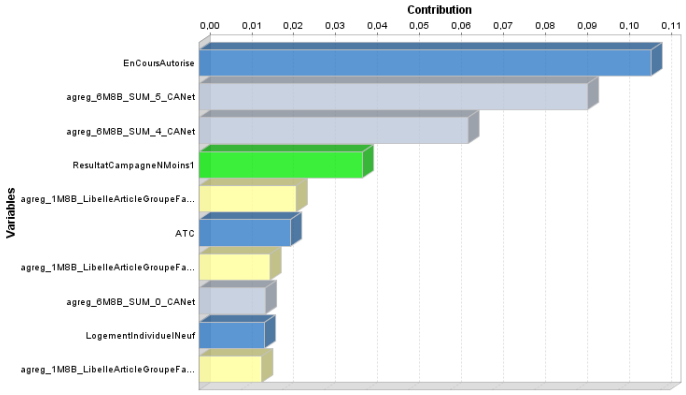


FIGURE 6.4 : Les dix variables les plus contributrices du modèle binaire

Les figures 6.5 et 6.6 illustrent le sens et la signification des valeurs des variables les plus contributrices. L'axe des ordonnées indique l'influence positive ou négative sur la variable binaire cible.

- La Figure 6.5 illustre la signification de la variable « chiffre d'affaires deux mois avant la campagne » (*Aggreg_6M8B_SUM_5_CANet* sur la Figure. 6.4). La majorité des clients ayant réalisé un chiffre d'affaires significatif deux mois avant la campagne (supérieur à 1 318 €) aurait tendance à participer à la campagne. Plus le chiffre d'affaires est important et plus la variable contribue positivement à la cible binaire.

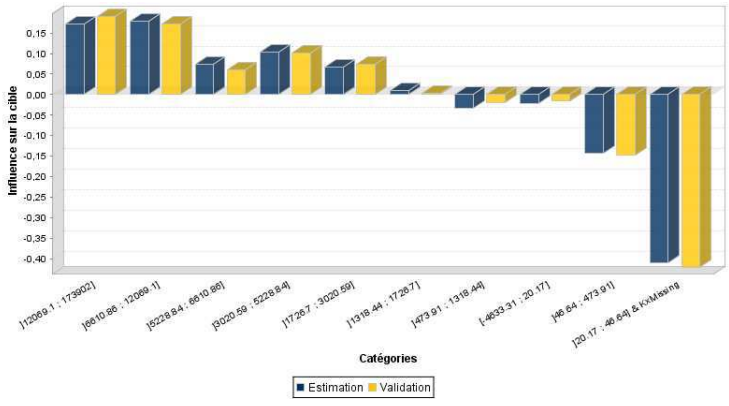


FIGURE 6.5 : Signification de la variable « chiffre d'affaires deux mois avant la campagne » pour le modèle à cible binaire

- La Figure 6.6 illustre la signification de la variable « ATC » (Attaché Technico-Commercial ou commercial) (cf. figure 6.4). Certains commerciaux contribuent positivement à l'achat durant la campagne marketing (*056F*, *083G*, etc.). A contrario, environ 35 % des commerciaux contribuent négativement à l'achat durant la campagne. Suite à ce constat, nous avons décidé d'enrichir les variables de la table « ATC » (cf. annexe A) dans l'entrepôt de données : taille du portefeuille client, proportion de clients actifs, nombre de visites mensuelles, prime sur objectif, etc. dans l'objectif de mieux comprendre le comportement des commerciaux (cf. concept « *Loop-Closed Mining* » définies dans la méthodologie *DDID-PD* présentée dans la section 2.2.2.2). La direction Négoce a réfléchi par la suite au rééquilibrage et au roulement des portefeuilles de ses commerciaux.

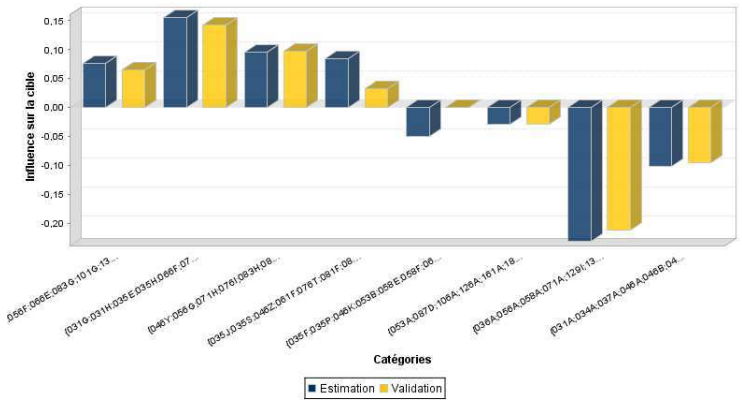


FIGURE 6.6 : Signification de la variable « ATC » pour le modèle à cible binaire

6.1.6 Application

L'application du modèle est réalisée grâce au jeu de données d'application, composé de 19 000 clients et présentant la même structure que les jeux de données binaire et continu. Pour les deux modèles, nous obtenons pour chaque client i deux variables p_i (probabilité de participation) et g_i (gain espéré de marge nette) permettant de trier les listes de routage des clients professionnels. Par exemple, dans le tableau 6.4, les clients *PR032912*, *PR032855*, *PR033033* et *PR159582* possèdent la même probabilité p_i (à 10^{-4} prêt) de participer à la campagne. Cependant, les clients *PR032912* et *PR159582* se voient monter de quelques rangs dans la liste car le gain espéré de marge nette g_i est deux fois plus important. La combinaison des deux modèles permet de répondre aux deux objectifs de la campagne : augmenter les ventes et améliorer le ROI. Générons la matrice des coûts dans le tableau 6.5, personnalisant ainsi la courbe de profit naïve (cf. figure 6.7). Pour accomplir cette tâche, nous définissons avec les experts métier la marge nette moyenne G réalisée par les clients durant la dernière campagne : environ 1 500 €. De la même manière, nous devons calculer la dépense moyenne de communication H par client : environ 62,5 €.

Identifiant client	**Probabilité p_i**	**Gain espéré g_i**	$p_i \times g_i$
PR032633	0,96501261	6 677,30	6 443,68
PR032785	0,965012363	5 162,11	4 981,50
PR032912	**0,964995889**	**4 384,68**	**4 231,20**
PR159582	**0,964965548**	**4 121,34**	**3 976,95**
PR032855	**0,964971733**	**2 240,29**	**2 161,82**
PR033033	**0,964968251**	**2 162,17**	**2 086,42**
PR033060	0,838573684	4 612,37	3 867,82
PR032857	0,793150561	2 281,12	1 809,27
PR167200	0,739359691	5 050,77	3 734,34
PR032996	0,588367688	1 504,34	885,11
PR032910	0,461373204	1 503,88	693,85
PR173542	0,289914103	1 112,60	322,56
PR160874	0,109833741	668,08	73,38
PR151560	0,081162517	662,77	53,79
PR033174	0,057455947	369,79	21,25

TABLEAU 6.4 : Données de base d'une liste de routage envoyée au commercial

		Réel	
		Acheteur	Non acheteur
Prédit	Acheteur (score $\geq s$)	$1500 - 62.5$	$0 - 62.5$
	Non Acheteur (score $< s$)	0	0

TABLEAU 6.5 : Matrice des coûts de la campagne (avec s un seuil de score)

Le maximum de la courbe naïve de profit (cf. figure 6.7) indique la proportion optimale de clients à contacter : 50.4 % correspondant à 88.14 % du profit maximum. Dès lors, notre liste de routage est composée de 50,4 % des 19 000 clients du jeu d'application, soit **9 575** clients ciblés.

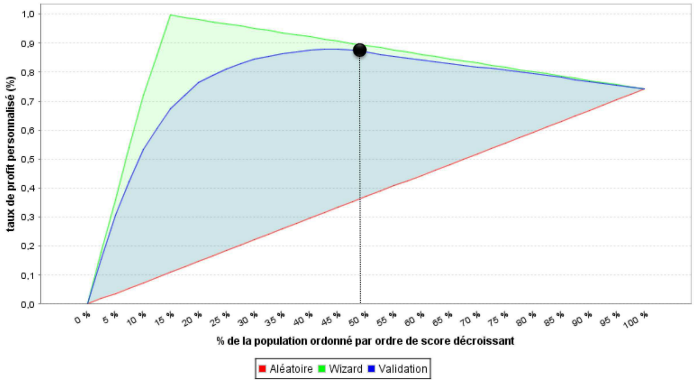


FIGURE 6.7 : Courbe de profit naïve sur le modèle binaire

6.1.7 Évaluation des résultats

La validation statistique des modèles a été contrôlée à l'aide des deux indicateurs KI et KR (cf. section 6.1.4). Cependant, les listes de routage « classiques » de l'équipe force de vente doivent être comparées aux liste de routage générées avec notre méthodologie. Environ 8 000 clients sont communs aux deux listes. De plus, tous les clients présents dans la liste de routage de l'équipe force de vente sont dans notre liste. Cependant, 1 575 clients ne sont pas inclus dans la liste de l'équipe force de vente. Ces derniers présentent des comportements atypiques : des clients très récents acheteurs (prospects) ou des clients en développement constant de chiffre d'affaires mais non suffisant pour être routé et quelques erreurs de fiabilité des données (métier ou nombre de salariés non renseigné, etc.) faisant que certains clients n'étaient pas routés.

6.1.8 Attribution des canaux de communication

L'expérience acquise sur dix années de routage clients lors de campagnes nous a conduit à définir trois classes de canaux de communication (cf. section 4.6.1 de notre méthodologie) représentant un choix stratégique pour VM Matériaux :

- **Classe N° 1** : {Visite, Courrier, Téléphone, SMS} ;
- **Classe N° 2** : {Courrier, Téléphone, Fax} ;
- **Classe N° 3** : {SMS, Email}.

La somme des coûts variables VC_j par classe du canal C (cf. tableau 6.6), a été calculée en utilisant le tableau 4.14 introduit dans la méthodologie. De plus, les valeurs d'atteinte At_j, et de conviction Cv_j des clients pour chaque canal j sont extraites du tableau 4.15, et les valeurs de α, β et γ sont exprimées dans le tableau 6.7 et validées par l'équipe force de vente de VM Matériaux.
L'affectation d'un client i à la meilleure classe de canal C_i est obtenue grâce à l'équation 4.16. Les résultats sont exprimés dans le tableau 6.8. Notons que l'affectation des clients aux canaux est limitée par la proportion maximale Ca_j (cf. tableau 4.15) de clients pouvant être contactés par le canal j en utilisant l'ensemble des ressources et données existantes.

Classe de canaux	Somme des coûts variables VC_j	Somme des coûts fixes FC_j
Classe N° 1	253.89 €	3 000 €
Classe N° 2	3.838 €	4 000 €
Classe N° 3	0.09 €	1 000 €

TABLEAU 6.6 : Sommes des coûts variables et fixes pour les classes de canaux

Pour répondre aux besoins de VM Matériaux, une contrainte sous forme de coefficient μ_{ij} est introduite dans l'équation 4.15 multipliant les coûts variables VC_j. En effet, μ_{ij} est égal à 1 excepté pour le canal j correspondant à la visite d'un commercial,

pour laquelle la valeur (supérieure à 1) dépend du chiffre d'affaires (ou du potentiel) du client i. L'objectif étant d'encourager la force commerciale à démarcher les clients présentant un fort potentiel mais un faible taux de pénétration [3] chez VM Matériaux. Ce coefficient concerne uniquement le canal *Visite* de la force commerciale.

Variables	Valeurs	Justification
α	50 000 €	coûts fixes de l'opération
β	100 000 €	licence et charge de travail du *data miner*
γ	5 000 € (5 jours-homme)	gain de temps pour le marketing

TABLEAU 6.7 : Valeurs des variables α, β et γ définies par les experts métier

Classe de canaux	Nombre de clients	Coûts (Section 4.6.1, Équation 4.14)
Classe N° 1	2 436	3 000 + 253.89 * 2 436 = 621 476,04 €
Classe N° 2	4 948	4 000 + 3.838 * 4 948 = 22 990,42 €
Classe N° 3	2 191	1 000 + 0.09 * 2 191 = 1 197,19 €
	9 575	645 663,65 €

TABLEAU 6.8 : Affectation des clients aux classes de canaux de communication

6.1.9 Retour sur investissement

Nous appliquons la formule 6.1 de la section 4.6.2 de notre méthodologie pour calculer le ROI espéré (pour des raisons de confidentialité, la partie gauche de la formule n'est pas instanciée). Pour chaque client i, l'affectation à la meilleure classe de canaux de communication C_i est arbitrée (cf. tableau 6.8).

$$ROI = \sum_i ProfitClient(i, C_i) - 8000 - (50\,000 + 100\,000 - 5\,000) \qquad (6.1)$$

L'estimation du profit a été très pertinente. En effet, le résultat effectif s'est avéré être l'estimation du profit (avec une erreur de marge inférieure à 5 %). Lors de la précédente opération commerciale, le taux de participation était de 18 %. Malgré la conjoncture qui s'est matérialisée par une baisse du CA sur les clients habituellement routés, la méthode a permis de sauvegarder le CA de l'opération commerciale. Au cours de la campagne ciblée, le taux de participation a atteint les **22 %** et **115** nouveaux clients présents dans notre liste de routage mais non dans celle du marketing ont participé, représentant environ un chiffre d'affaires additionnel de **1 200 000** €.

6.1.10 Généralisation et automatisation

Cette première expérience sur une campagne marketing nous a permis d'étendre l'utilisation de notre méthodologie à toutes les campagnes marketing *one-to-one* [4] du

3. Le taux de pénétration correspond à la masse marge réelle du client divisé par la masse marge potentielle du client sur une année.

4. Le marketing *one-to-one* est un marketing individualisé, par opposition au marketing de masse.

groupe VM Matériaux. La construction des trois jeux de données (cf. tableau 6.2) et la génération des modèles binaire et continu ainsi que leur application ont été automatisés et planifiés. Les résultats des modèles (probabilité et gain espéré durant la campagne) sont automatiquement intégrés dans l'entrepôt de données (cf. tables *Campagne*, *CampagneRoutage* et *CampagneResultat* en annexe A). Les résultats servent ainsi de base à des restitutions sous forme de *reporting*, à des cubes multidimensionnels ou à d'autres modèles répondant à des problématiques métier différentes (cf. figure 6.8).

Client	Nom Client	Ctr	ATC	Palier1	Palier2	Eg	Sl	Ba	CA à fin 05/2011	Mg à fin 05/2011	CA à fin 05/2010	Mg à fin 05/2010	% P1	%P2	Avert.	Ev	Mtt place(s) jouée(s)	Résultat
Région																		
Agence																		
PR028696		DEV	066H -	26 000	31 500	2			3 942	650	2 536	618	15,16 %	12,52 %			1 150	
PR028777		DEV	066F -	55 800	71 200		2		4 632	-270	17 414	2 366	8,30 %	6,51 %			2 000	
PR028869		DEV	066F -	145 500	162 200		2		112 147	7 200	86 372	11 345	77,08 %	69,14 %			2 000	
PR028899		DEV	066F -	249 303	265 731	4			181 399	26 673	126 373	18 607	72,76 %	68,26 %			2 300	
PR028997		DEV	066F -	249 300	271 800	2	2		205 708	30 523	120 168	18 160	82,51 %	75,68 %			3 150	
PR029183		FID	066E -	360 000	1,00 %	2	2	2	391 331	54 703	307 169	44 188	108,70 %				6 050	
PR029246		DEV	066F -	108 000	125 000		2		78 066	8 938	58 595	6 716	72,28 %	62,45 %			2 000	
PR029263		FID	066E -	295 000	1,00 %		2	2	208 149	35 772	185 198	30 763	70,56 %				4 900	
PR029387		DEV	066E -	58 800	74 200		2		102 029	12 871	20 349	2 618	173,52 %	137,51 %			2 000	100 % place(s) gagnée(s)
PR150258		DEV	066E -	158 400	175 000	2	2		136 954	24 238	62 914	12 227	86,46 %	78,26 %			3 150	
PR152525		DEV	066E -	167 200	175 400	2			66 647	12 240	136 747	19 323	39,86 %	38,00 %			1 150	
PR155068		DEV	066F -	100 700	112 200	1			114 917	11 303	50 742	5 515	114,12 %	102,42 %			575	100 % place(s) gagnée(s)
PR155837		DEV	066F -	172 890	193 330	1		2	136 455	25 295	94 150	15 543	78,93 %	70,58 %			3 475	
PR155996		DEV	066E -	54 500	63 300	2			40 750	3 461	7 294	1 214	74,77 %	64,38 %			1 150	
PR176256		DEV	066F -	40 600	53 400	2			35 103	6 365	3 476	326	86,46 %	65,74 %			1 150	
PR178572		DEV	066H -	25 455	30 682	2			4 731	898	14 514	3 619	18,59 %	15,42 %			1 150	
PR288689		DEV	066E -	26 000	31 500	2			15 945	2 831	0	0	61,33 %	50,62 %			1 150	
Total				2 293 448	1 836 445	24	16	6	1 838 905	263 691	1 294 011	193 149	80,18 %	100,13 %			38 500	
Total				**2 293 448**	**1 836 445**	**24**	**16**	**6**	**1 838 905**	**263 691**	**1 294 011**	**193 149**	**80,18 %**	**100,13 %**			**38 500**	
Total Général		**17**		**2 293 448**	**1 836 445**	**24**	**16**	**6**	**1 838 905**	**263 691**	**1 294 011**	**193 149**	**80,18 %**	**100,13 %**			**38 500**	

FIGURE 6.8 : Un modèle de risque sur objectif durant une autre campagne marketing

De plus, certaines déviations du modèle peuvent être détectées d'une année sur l'autre, rendant ainsi nécessaire la régénération du modèle avant sa simple application. Par exemple, le comportement des clients professionnels à la fin de l'année 2009 s'est vu impacté par la crise financière.

Actuellement, nos modèles de fouille de données sont « industrialisés » [94] et les résultats sont stockés dans l'entrepôt de données avec un flux de validation réalisé par l'équipe force de vente. Les modèles ont été généralisés aux trois principales campagnes marketing du groupe en Avril, Septembre et Novembre de chaque année (cf. tableau 6.9) : l'évolution de la participation correspond au nombre de clients participants ayant dépassé un seuil prédéfini de chiffre d'affaires, l'évolution des ventes compare les résultats à la campagne marketing similaire de l'année précédente.

Campagnes	Période	Évolution de la participation	Évolution des ventes
1	Avril	+4 %	+5 %
2	Septembre	+7 %	+3 %
3	Novembre	+2 %	équivalent

TABLEAU 6.9 : Industrialisation des modèles aux campagnes de VM Matériaux

Faisant référence à la section 4.6.2, la généralisation de nos modèles sur trois campagnes marketing a permis de réduire le coût moyen de chaque campagne, c'est-à-dire les coûts fixes α de 50 000 à 16 500 € et les coûts de la fouille de données β de 100 000 à 33 000 € par campagne.

6.2 Méthodologie de recommandations actionnables et profitables

L'activité Négoce du groupe VM Matériaux [5] propose à ses clients professionnels plus de 100 000 références produits à travers 116 magasins. Les clients artisans du bâtiment présentent des comportements d'achats différents et évolutifs ainsi que des variables descriptives souvent atypiques. Dès lors, l'utilisation de notre méthodologie *CAPRE* s'avère pertinente pour recommander des produits à forte valeur ajoutée à des clients susceptibles d'en acheter et ainsi créer une relation personnalisée et privilégiée avec le client.

6.2.1 Préparation des données

Pour remédier au problème de *sur-spécialisation* [16], nous utilisons les sept niveaux de la taxonomie produits de VM Matériaux (cf. figure 6.9). Le niveau le plus fin permet d'entreprendre des actions commerciales plus précises, mais multiplie les règles. Travailler au niveau le plus général permet d'obtenir des règles plus fortes. Cette façon de procéder permet d'obtenir des règles plus pertinentes, dans lesquelles les produits les plus courants ne dissimulent pas, par leur fréquence, les produits les moins courants.

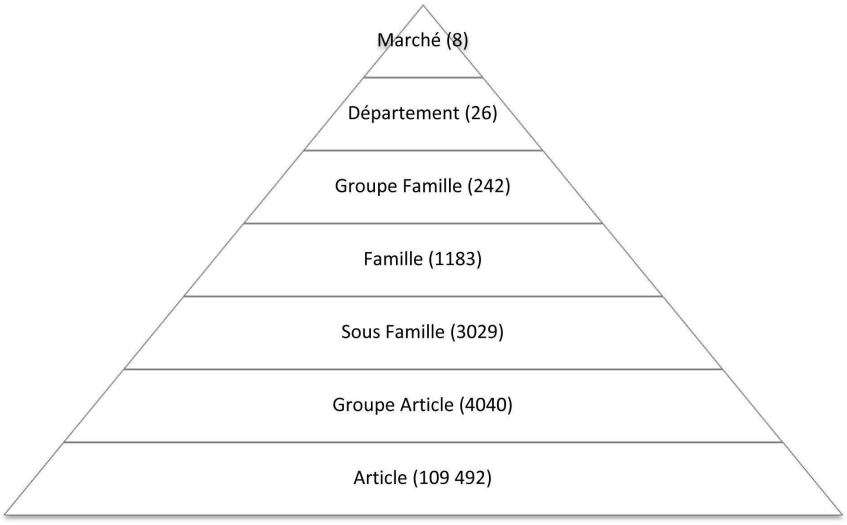


FIGURE 6.9 : Taxonomie produits de VM Matériaux

EXEMPLE 11 La plaque de plâtre standard de 13 mm, de 250 $\times$ 120 cm possède la hiérarchie suivante (du groupe article au marché) : plaque et plâtre standard, plaque standard, plaque de plâtre standard, plaque de plâtre, cloison et plafond, isolation.

5. http://www.vm-materiaux.fr/activites/negoce-materiaux

EXEMPLE 12 Les produits les plus rares tels que les panneaux photovoltaïques sont codifiés à un niveau plus général, tandis que les produits les plus courants comme la plaque de plâtre standard de 13 mm sont codifiés à un niveau plus fin.

À l'aide des experts métier de VM Matériaux, nous avons décidé de regrouper les items sur trois niveaux de la hiérarchie produits (cf. figure 6.10) les plus représentatifs et compréhensibles pour les métiers : le groupe famille, la famille et la sous famille de produits. Nous avons également fixé un paramètre de fréquence maximale d'achat des produits $Fq_{max} = 2\,000$, supprimant ainsi les groupes de produits trop fréquents, c'est-à-dire dont la fréquence d'achat est supérieure à 2 000 occurrences sur une période prédéfinie. Par exemple, l'algorithme[6] se déroule comme suit (cf. figure 6.10) :

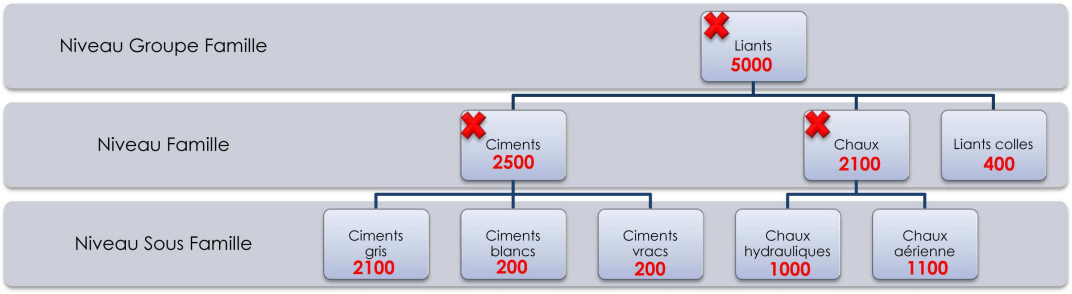


FIGURE 6.10 : Parcours de la taxonomie produits avec $Fq_{max} = 2\,000$

- Le groupe famille des *Liants* ayant une fréquence de 5 000 (supérieure à Fq_{max}) est trop fréquent. Dès lors, nous descendons d'un niveau dans la hiérarchie produits et supprimons le groupe famille des *Liants*.
- Les familles *Ciments* et *Chaux* ayant des fréquences respectives de 2 500 et de 2 100 sont trop fréquentes. Dès lors, nous descendons d'un niveau dans la hiérarchie produits et supprimons les familles des *Ciments* et *Chaux*. En revanche, la fréquence de la famille *Liants Colles* est inférieure à Fq_{max}. Par conséquent, cette famille de produits est conservée et les sous-items ne sont pas parcourus.
- Enfin, au dernier niveau, toutes les sous familles dépendantes d'une famille non supprimée sont conservées.

Cet exemple permet de réduire le nombre d'items de la branche *Liants Colles*. Le niveau le plus fin de la taxonomie (l'article) n'est jamais directement recommandé, permettant ainsi de laisser un libre choix au client de choisir sa marque et ses préférences. Les 100 000 produits actifs ont ainsi été regroupés en 431 taxons (P) (groupe famille, famille ou sous familles) sur lesquels la méthodologie *CAPRE* est appliquée.

6. L'algorithme a été réalisé sous forme de procédure SQL directement dans le SGBD de VM Matériaux. Le nombre de niveaux parcourus dans la taxonomie et la fréquence Fq_{max} sont des paramètres de la procédure.

Les 9 575 clients (C) ciblés dans la phase préparatoire de la section 6.1 sont ensuite décrits à travers 461 variables (cf. tableau 6.11) divisées en trois catégories :

- **21 variables internes** (dans l'entrepôt de données) : adresse, métier, fidélité, type de profil (mouvementé, endormi, prospect), adhérence (carte de fidélité), ancienneté, commercial, magasin de rattachement, encours autorisé, etc.
- **9 variables externes** (généralement obtenues à l'aide d'un fichier *Coface*) : nombre d'employés, catégorie socio-professionnelle, bilan, etc.
- **431 agrégats de chiffre d'affaires** : calculés dans le SGBD pour chaque client à travers la taxonomie produits.

En accord avec les experts métier, les **288 008** valeurs des variables de chiffre d'affaires $u(p,c)$ sont rendues commensurables ($u^{\%}(p,c)$) et sont discrétisées ($u^*(p,c)$) en trois intervalles a, b et c de même fréquence $\frac{1}{3}$, identifiant trois niveaux de part de chiffre d'affaires : faible, moyen et fort.

EXEMPLE 13 Dans le tableau 6.10, nous présentons les deux transformations appliquées sur les chiffres d'affaires $u(p,c)$ de trois carreleurs sur la famille de produits Carrelage. Dans un premier temps, la part de chiffre d'affaires de carrelage $u^{\%}(p,c)$ est calculée par rapport au chiffre d'affaires global du client (*CA Total*). L'étape de discrétisation $u^*(p,c)$ permet ensuite d'analyser que, bien que le client *PR175058* (18 559 €) ait acheté moins de carrelage que les clients *PR136671* et *PR048123*, il représente par rapport à son chiffre d'affaires total le plus « fort » acheteur de carrelage avec $u^*(p,c) = c$.

Client	Métier	Nb Salariés	CA Total	$\mathbf{u(p,c)}$	$\mathbf{u^{\%}(p,c)}$	$\mathbf{u^*(p,c)}$
PR136671	Carreleur	40	232 859 €	122 231 €	52,49 %	a
PR048123	Carreleur	7	55 205 €	38 548 €	69,83 %	b
PR175058	Carreleur	1	20 779 €	18 559 €	89,32 %	c

TABLEAU 6.10 : Transformation des variables d'achats $u(p,c)$ de trois clients

Cette phase de sélection et de transformation des variables permet d'obtenir 461 variables. Le tableau 6.11 récapitule la cardinalité de chaque variable utilisée pour l'application de la méthodologie *CAPRE*.

Variable	Nombre	Description
$\lvert C\rvert$	9 575	clients ciblés lors de la phase préparatoire (cf. section 6.1)
$\lvert P\rvert$	431	regroupement des 100 000 items en 431 agrégats de CA
$\lvert N\rvert$	22	variables descriptives numériques des clients
$\lvert I\rvert$	8	variables descriptives catégoriques des clients
$u(p,c)$	288 008	les chiffres d'affaires du produit p pour le client c

TABLEAU 6.11 : Cardinalité des variables utilisées dans l'application de *CAPRE*

6.2.2 Extraction des cohortes de règles

L'extraction des règles est réalisée à l'aide de l'algorithme *CHARM*[7] (cf. section 5.1.2.3) fondé sur la recherche en profondeur d'itemsets fermés [306]. Avec un support de 40 clients et une confiance de 50 %, **23 578** règles sont extraites (cf. tableau 6.12).

Support (%)	Confiance (%)	Lift	Règles
0.51	53,45	11,91	SF_COLLE EN PATE = c & F_CROISILLONS = c → F_OUTILS CARRELAGE = c
0.74	61,38	24,52	SF_LITEAUX SAPIN = c & SF_TUILES ROMANES = c → SF_ACC TUILES ROMANES = c
0.64	71,96	12,60	SF_CIMENTS GRIS = c & SF_GRAVIERS RIVIERE = c → SF_SABLE DE CARRIERE = c
0.49	57,28	22,88	SF_CIMENTS GRIS = b & SF_TUILES ROMANES = c → SF_ACC TUILES ROMANES = c
0.64	53,42	11,91	SF_MORTIERS COLLES = c & SF_TALOCHE_SCEAU = c → F_OUTILS CARRELAGE = c

TABLEAU 6.12 : Exemples de règles d'association extraites par l'algorithme *CHARM*

Le *lift* (cf. figure 6.11) sur l'ensemble des règles varie entre 2,87 et 51,48 avec une moyenne de 9,71, ce qui dénote des attractions fortes entre les achats de produits.

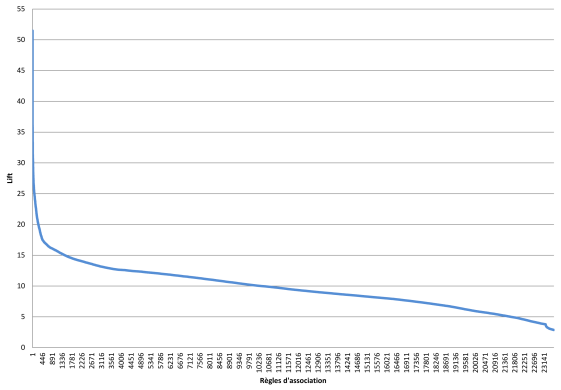


FIGURE 6.11 : Évolution du *lift* sur les 23 578 règles d'association

Les 23 578 règles permettent de générer un ensemble de **174** cohortes. Le tableau 6.13 présente des exemples de cohortes sélectionnées par les experts métier ainsi que leurs caractéristiques : nombre d'exemples (# Ex), nombre de contre-exemples (# CE), *lift* moyen et conclusion.

N° cohorte	# Ex	# CE	Lift moyen	Conclusion
4	98	96	22,68	GF_EVIERS MEUBLES=c
20	96	64	32,17	F_DALLES CLASSIQUES=c
22	165	83	20,35	GF_PLATRES=c
23	151	88	28,57	GF_EQUIPEMENT SANITAIRE=c
58	63	53	10,71	SF_SEMELLES LONGRINES=b
61	242	209	16,96	F_TUILES ROMANES=c

TABLEAU 6.13 : Exemples de cohortes générées à partir des 23 578 règles

7. Codé en C++ par son concepteur Mohammed J. ZAKI, http://www.cs.rpi.edu/~zaki/software/

Pour la suite de l'expérience, nous concentrons la présentation des résultats de la méthodologie sur la cohorte N° 22 recommandant l'achat du groupe famille des plâtres (*GF_PLATRES = c*), composée des plâtres à projeter, des plâtres manuels traditionnels ou allégés ainsi que des plâtres génériques. La cohorte est composée de 14 règles d'association (cf. tableau 6.14) concluant sur le même groupe famille *GF_PLATRES = c*. Les experts métier affirment que l'ensemble des prémisses des 14 règles constituent des achats correspondant aux règles métier des plâtriers ou plaquistes, e.g. les cloisons, les plaques de plâtre, les enduits, les rails et montants métalliques, etc.

Sup.(%)	Conf.(%)	Lift	Règle
1,63	69,12	17,74	GF_CLOISONS PLAFONDS BRIQUES=c → GF_PLATRES=c
0,55	77,14	19,82	F_COLLES ENDUITS CARREAUX PLATRE=c → GF_PLATRES=c
0,70	75,52	19,43	SF_PLAQUE STANDARD=c & GF_CLOISONS PLAFONDS BRIQUES=c → GF_PLATRES=c
0,62	84,03	21,75	SF_LAINE VERRE PANNEAUX PAROIS=c & GF_CLOISONS PLAFONDS BRIQUES=c → GF_PLATRES=c
0,57	73,68	18,93	GF_CLOISONS PLAFONDS BRIQUES=c & SF_TRANSPORT MATERIAUX=b → GF_PLATRES=c
0,51	79,77	20,40	GF_CLOISONS PLAFONDS BRIQUES=c & SF_RAILS ET MONTANTS=b → GF_PLATRES=c
0,56	83,28	21,44	GF_CLOISONS PLAFONDS BRIQUES=c & SF_FOURRURES=c → GF_PLATRES=c
0,52	89,13	23,04	GF_CLOISONS PLAFONDS BRIQUES=c & SF_VISSERIE POINTES PLAQUES DE PLAT=c → GF_PLATRES=c
0,91	81,10	20,79	GF_CLOISONS PLAFONDS BRIQUES=c & SF_COLLES ENDUITS=c → GF_PLATRES=c
0,52	88,89	22,64	GF_CLOISONS PLAFONDS BRIQUES=c & SF_PLAQUE STANDARD=b → GF_PLATRES=c
0,63	79,12	20,38	GF_CLOISONS PLAFONDS BRIQUES=c & SF_RAILS MONTANTS=c → GF_PLATRES=c
0,50	91,02	23,37	GF_CLOISONS PLAFONDS BRIQUES=c & SF_RUBANS ADHESIFS=c → GF_PLATRES=c
0,56	78,09	20,20	SF_PLAQUE STANDARD=c & GF_CLOISONS PLAFONDS BRIQUES=c & SF_COLLES ENDUITS=c → GF_PLATRES=c
0,51	79,24	20,40	GF_CLOISONS PLAFONDS BRIQUES=c & SF_COLLES ENDUITS=c & SF_RAILS MONTANTS=c → GF_PLATRES=c

TABLEAU 6.14 : Les 14 règles d'association de la cohorte $Ct(GF\_PLATRES = c)$

L'agrégation des règles en une cohorte *Ct(GF_PLATRES = c)* génère un ensemble de **165** exemples $Ct^+(GF\_PLATRES = c)$ et **83** contre-exemples $Ct^-(GF\_PLATRES = c)$. Le tableau 6.15 présente quelques caractéristiques descriptives des exemples et contre-exemples de la cohorte.

	Exemples $\mathbf{Ct^+(GF\_PLATRES = c)}$	**Contre-exemples** $\mathbf{Ct^-(GF\_PLATRES = c)}$
Nombre de clients	165	83
Métier	72 % plâtriers/plaquistes	25 % plâtriers/plaquistes
Nombre moyen de salariés	$\simeq 5$	$\simeq 7$
Nombre de carte de fidélité	102	61
CA net annuel moyen	56 000 €	63 000 €
Part de CA d'isolation moyenne	70,30 %	36,92 %

TABLEAU 6.15 : Description des $Ct^+(GF\_PLATRES = c)$ et $Ct^-(GF\_PLATRES = c)$

La partie suivante présente les étapes de pré-actionnabilité et d'actionnabilité permettant d'élaguer les 83 contre-exemples de la cohorte $Ct(GF\_PLATRES = c)$.

6.2.3 Mesure de l'actionnabilité des contre-exemples

6.2.3.1 Pré-actionnabilité sur les variables d'achats

Visualisons tout d'abord les distributions des exemples Ct^+ et contre-exemples Ct^- sur les achats des prémisses d'une des règles de la cohorte (cf. figure 6.12). Utilisons la règle : SF_PLAQUE STANDARD=c & GF_CLOISONS PLAFONDS BRIQUES=c & SF_COLLES ENDUITS=c → GF_PLATRES=c.

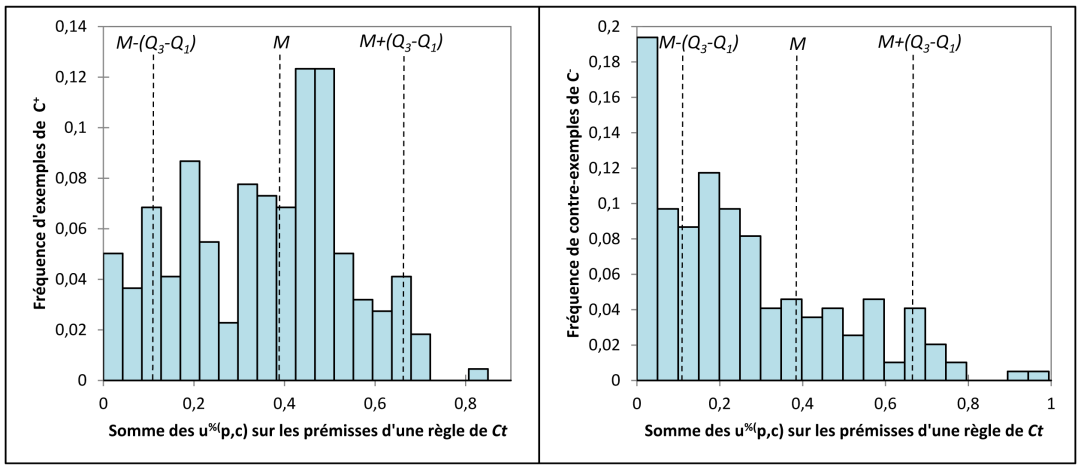


FIGURE 6.12 : Distribution des achats en prémisses des Ct^+ et Ct^- pour la règle SF_PLAQUE STANDARD=c & GF_CLOISONS PLAFONDS BRIQUES=c & SF_COLLES ENDUITS=c → GF_PLATRES=c

Les achats en prémisse des Ct^+ de la règle de la cohorte sont caractérisés par une médiane $M = 26,02$ % et un intervalle interquartile $Q3 - Q1 = 14,05$ %. Ces derniers sont reportés sur la distribution des Ct^-. Par conséquent, les Ct^- en dehors de l'intervalle $[11,97\ \%;\ 40,06\ \%]$ sont considérés comme non pré-actionnables. Cette manipulation est réalisée sur l'ensemble des règles de la cohorte. Ce filtrage permet d'élaguer **5** contre-exemples. À travers cet élagage, les experts métier identifient deux comportements d'achats sur les prémisses des règles de la cohorte :

1. Soit le client présente trop peu de dépense sur les prémisses de l'ensemble des règles. Dès lors, selon les experts métier, il s'agit bien de dépannage pour l'artisan qui habituellement se fournit à la concurrence ;

2. Soit le client présente la quasi totalité de ses dépenses sur les prémisses des règles, dès lors les experts estiment qu'il existe une raison métier (concurrence, tarif, clients grands comptes tels que *Bouygues* ou *Eiffage*, etc.) justifiant le non achat de la conclusion de la cohorte à hauteur du chiffre d'affaires espéré.

Dès lors, nous nous intéressons aux **78** contre-exemples pré-actionnables.

6.2.3.2 Actionnabilité sur les variables descriptives

Afin de sélectionner les contre-exemples actionnables de *Ct(GF_PLATRES=c)*, nous nous intéressons à la matrice des distances sur les variables descriptives. La distance moyenne sur l'ensemble de la matrice des distances est $minDist = 28,02$. Nous définissons avec la direction commerciale le seuil de part de voisinage $\delta = 10\,\%$. Dès lors, un contre-exemple pré-actionnable est actionnable si et seulement si il possède 10 % des exemples de Ct^+ dans son voisinage à une distance inférieure à $28,02$. Parmi les 78 contre-exemples pré-actionnables, **5** contre-exemples (cf. tableau 6.16) possèdent moins de 10 % des exemples (moins de 21 clients exemples) à une distance inférieure ou égale à $28,02$ dans l'espace des variables descriptives.

Identifiant client	Nombre de salariés	Carte de fidélité	CA net annuel	Métier	Part des achats pour l'isolation
PR161080	16	oui	442 450 €	Plaquiste	95,43 %
PR035021	15	oui	278 500 €	Plaquites	98,70 %
PR025566	45	oui	577 600 €	Plaquiste	91,85 %
PR032047	49	oui	698 400 €	Plâtrier	94,19 %
PR160087	5	non	8 500 €	Négociant	6,82 %

TABLEAU 6.16 : Les cinq contre-exemples non actionnables élagués

Les experts métier identifient un client négociant (*PR160087*) présentant des valeurs de variables descriptives éloignées. Les quatre autres clients sont des plaquistes et plâtriers importants. Ces clients achètent généralement chez plusieurs négociants de matériaux pour obtenir le meilleur prix. Ce deuxième filtrage permet de sélectionner **73** clients actionnables.

6.2.4 Mesure de l'intérêt économique des cohortes

6.2.4.1 Profitabilité a priori

À l'aide du directeur marketing de VM Matériaux, nous avons considéré que les coûts variables d'une visite d'un client par un commercial étaient de 250 € (cf. tableau 4.14 de notre méthodologie). Dès lors, nous avons pu fixer le seuil $\theta = 1000$ € [8] à partir duquel un contre-exemple est profitable pour l'entreprise. En appliquant la formule 4.13 de notre méthodologie, **59** contre-exemples actionnables sont 1000 €-profitables, c'est-à-dire que leur dépense espérée pour l'achat du groupe famille $GF\_PLATRES = c$ est supérieure à 1 000 € de chiffre d'affaires annuel. Un récapitulatif du filtrage des contre-exemples de la cohorte est illustré en figure 6.13.

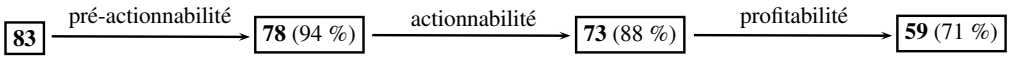


FIGURE 6.13 : Filtrage des contre-exemples de la cohorte $Ct(GF\_PLATRES = c)$

8. Le seuil θ a été fixé par les experts métier considérant la marge nette minimale à dégager pour que le démarchage d'un client soit profitable pour l'entreprise.

Pour les 59 contre-exemples profitables, l'intérêt économique espéré pour la cohorte de règles est estimé à environ 568 980 € de chiffre d'affaires annuel sur le groupe famille des plâtres, soit environ 9 600 € par client. En appliquant la formule 4.17 de notre méthodologie, nous démarchons l'ensemble des 59 clients à l'aide de la force commerciale (coût fixe de 250 € par démarchage client) et nous estimons une marge commerciale moyenne de 20 %. Dès lors, le ROI du déclenchement de la cohorte est estimé à 99 046 €.

6.2.4.2 Profitabilité personnalisée

Pour affiner la profitabilité de chaque client, nous décidons d'utiliser un *scoring* avec l'outil de fouille de données KXEN. Pour prédire la cible continue de chiffre d'affaires par client pour le conséquent *GF_PLATRES=c*, nous choisissons de calculer un score de tendance à l'aide d'une régression *ridge*. La phase d'apprentissage est guidée par les 165 exemples de la cohorte et la phase d'application par les 59 contre-exemples profitables. Le jeu de données d'apprentissage présente peu de clients. Ils représentent cependant une population très homogène de clients avec un fort potentiel d'achat sur le conséquent de la cohorte et des variables descriptives similaires, améliorant ainsi la précision du modèle. Pour les 59 clients profitables, le profit personnalisé pour la cohorte est estimé à environ 317 340 €, i.e environ 5 400 € par client. Dès lors, le ROI est estimé à 48 718 €.

6.2.5 Impact de la profitabilité sur les *Top-20 ROI* recommandations

Penchons-nous maintenant sur les *Top-20 ROI* recommandations triées par retour sur investissement et analysons l'impact de l'étape de profitabilité de la méthodologie *CAPRE* (cf. tableau 6.17).

Top-20 ROI	*nbCE*	Avant profitabilité			Après profitabilité		
		nbCEAct	ROI (€)	ROI moyen par client (€)	nbCEProfit	ROI (€)	ROI moyen par client (€)
Ciments gris	2 455	1 876	928 125	495	1 230	1 011 759	823
Blocs creux	1 441	1162	766 688	660	857	803 460	938
Bétons courants standard	249	213	452 486	2 134	205	453 272	2 211
Menuiseries extérieures	223	161	423 061	2 628	161	423 061	2 628
Plaque de plâtre standard	1 193	1 091	373 871	343	656	427 082	651
Enduits mono-couches	229	207	309 384	1 495	197	310 176	1 574
Bétons hors normes	65	54	257 677	4 772	54	257 677	4 772
Tuiles romanes	209	188	195 926	1 042	165	198 145	1 201
Palettes facturées	1 495	1 209	195 821	162	581	290 986	501
Bois de charpente sapins	444	372	154 666	416	267	168 188	630
Laine verre rouleaux	647	598	122 961	206	284	169 441	597
Mortiers colles	989	824	120 473	146	400	182 488	456
Laine verre panneaux	561	499	113 747	228	242	151 601	626
Treillis bâtiment	839	759	104 825	138	358	169 111	472
Plâtres	83	73	97 481	1 335	59	99 046	1 679
Rails et montants métalliques	1 142	1 041	56 456	54	390	158 728	407
Plaque de plâtre hydrofuge	901	839	-1 878	-2	259	91 348	353
Fers tort et ronds	802	719	-34 270	-48	226	54 754	242
Colles et enduits pour plaques	1 148	974	-48 867	-50	263	74 949	285
Fourrures	995	896	-48 654	-54	224	68 557	306
		13 755	4 539 979	805	7 078	5 563 829	1 067

TABLEAU 6.17 : Impact de la profitabilité sur les *Top-20 ROI* recommandations

Tout d'abord, le nombre de contre-exemples à démarcher diminue fortement avant et après la phase de profitabilité. En effet, le nombre de clients professionnels à recommander passe de 13 755 clients contre-exemples actionnables (*nbCEAct*) à 7 078 clients contre-exemples 1000 €-profitables (*nbCEProfit*). Dès lors, le retour sur investissement (*ROI*) est impacté à la hausse en passant de 4 539 979 € à 5 563 829 €. En effet, les coûts de démarchage sont amoindris par la diminution de la cible. Le ROI définit dans la méthodologie (cf. section 4.6.2) peut être simplifié comme suit :

$$ROI = (profitabilite \times 20\,\%) - (250 \times \textit{nombre de clients demarchables})$$

Également, le retour sur investissement moyen par client subit une hausse d'environ 20 % en passant de 805 € à 1 067 €.
Analysons plus en détail la recommandation de laine de verre en rouleaux (*Laine verre rouleaux*) dans le tableau 6.17. Le nombre de contre-exemples à démarcher passe de 598 à 284 clients 1000 €-profitables. Le retour sur investissement moyen par client est triplé et avoisine les 600 €. Le retour sur investissent de la recommandation suggérant les colles et les enduits pour plaque de plâtre (*Colles et enduits pour plaques*) dans le tableau 6.17 passe d'un ROI négatif de -48 867 € à un ROI profitable pour l'entreprise de 74 949 €.
Remarquons que, plus nous déclenchons d'actions sur les cohortes et les clients profitables et moins les charges fixes affectent le retour sur investissement.

6.2.6 Recommandations aux clients de VM Matériaux

Notre méthodologie a permis à l'équipe commerciale d'analyser le comportement de leurs clients de manière à comprendre les raisons pour lesquelles certains clients n'achètent pas certains produits (tarif, concurrence, etc.).

EXEMPLE 14 Un expert métier du Négoce de matériaux a validé que les recommandations proposées à un plaquiste à très fort potentiel d'achat étaient pertinentes. En effet, les artisans présentant un fort potentiel d'achats réservent souvent une « part » de leur chiffre d'affaires pour chaque fournisseur de matériaux, négociant ainsi en permanence les prix. Ces clients ont également tendance à stocker eux mêmes leurs matériaux pour améliorer leur négociation d'achats et la productivité de leurs salariés (gain de temps, trajet direct entre l'entreprise et le chantier).

Le nombre restreint de recommandations profitables permet d'afficher des indicateurs intelligibles dans un outil de CRM pour aider les commerciaux à développer la valeur de leurs clients. Les experts métier peuvent expliquer et justifier les recommandations à l'aide des prémisses des règles d'association des cohortes.
La figure 6.14 illustre la distribution du nombre de contre-exemples actionnables et de la profitabilité a priori en Euros de chaque cohorte. Trois types de cohortes (cas N° 1, 2 et 3 sur la figure 6.14) se distinguent :

Cas N° 1 les cohortes présentant une profitabilité importante mais un nombre de contre-exemples démarchables trop important ;

Cas N° 2 les cohortes présentant une profitabilité moyenne et un nombre de contre-exemples démarchables atteignables par la capacité de notre force de vente (cf. critère de capacité du tableau 4.15 de notre méthodologie) ;

Cas N° 3 les cohortes présentant une profitabilité très faible, voire nulle.

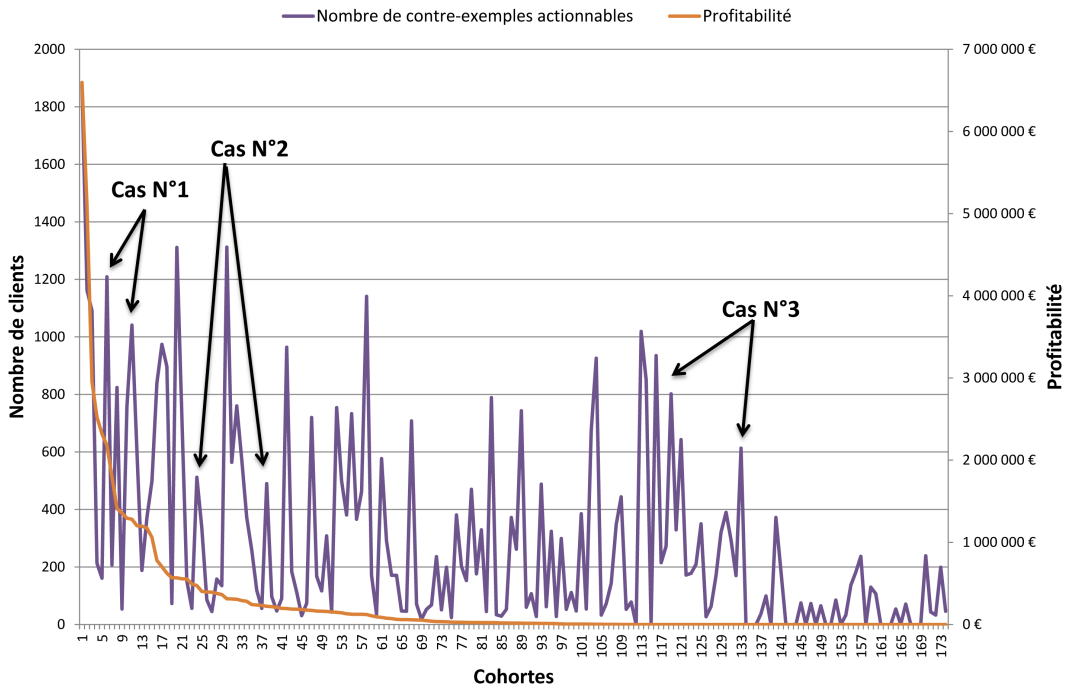


FIGURE 6.14 : Contre-exemples actionnables et profitabilité des cohortes

Des démarches davantage marketing peuvent être mises en place :

1. Promouvoir des produits sur des actions commerciales ciblées (produits gratuits dans le coffre des commerciaux) en sélectionnant les cohortes les plus profitables et leurs contre-exemples actionnables (cf. figure 6.14).

2. Cibler les clients d'une région ou d'un magasin et recommander trois à quatre produits en fonction des cohortes pour lesquelles des clients sont contre-exemples actionnables et intéressants économiquement.

Enfin, en fonction des recommandations économiquement intéressantes, des partenariats fournisseurs peuvent être convenus pour mettre en valeur des produits sur une période prédéfinie et dans des zones de chalandise ciblées.

6.2.7 Validation des résultats

6.2.7.1 Validation croisée

Le blanchiment des données

Afin d'évaluer la qualité de notre méthodologie de recommandations actionnables et profitables sur des données opérationnelles et volumineuses, nous réalisons une validation croisée sur les données de VM Matériaux. Rappelons que le jeu de données repose sur des achats réels sur l'année 2010 et fournit de ce fait un bon support de validation. En effet, la matrice contient 9 575 clients, 431 items et 288 008 achats. Ainsi, la matrice d'achats présente une dispersion de 6,98 % (cf. section 3.2.6.1), c'est-à-dire qu'il y a 93,02 % de données manquantes, considérées comme des non-achats ou des avoirs.
Le jeu de données a été blanchi [9] aléatoirement de 10 % des achats, c'est-à-dire 28 001 achats. Le taux de blanchiment est moins important que la validation sur le jeu de données *MovieLens*. En effet, les 431 variables d'achats du jeu de données VM Matériaux représentent des agrégats de chiffre d'affaires. Dès lors, chaque blanchiment peut correspondre à plusieurs achats d'un même client sur l'année.
Cinq ensembles d'apprentissage et cinq ensembles de test appelés respectivement « base » et « test » ont ainsi été créés. *VM.data* correspond au jeu de données complet. *VM[1-5].base* sont les cinq ensembles d'apprentissage et *VM[1-5].test* sont les cinq ensembles de validation générés. Les ensembles d'apprentissage et de test contiennent respectivement 90 % et 10 % des achats globaux. Dans le tableau 6.18, nous analysons l'impact du blanchiment des valeurs supérieures à 1 000 € et le nombre moyen de valeurs par client.

Caractéristiques / Jeux de données	**Nombre de CA**	**Nombre de CA > 1000 €**	**Nombre de CA moyen par client**
VM.data	288 008	42 784	30,03 %
VM1.base	259 207	38 478	27,07 %
VM2.base	259 207	38 383	27,06 %
VM3.base	259 207	38 519	27,07 %
VM4.base	259 207	38 540	27,07 %
VM5.base	259 207	38 495	27,07 %
VM1.test	28 801	4 306	2,93 %
VM2.test	28 801	4 401	2,94 %
VM3.test	28 801	4 265	2,93 %
VM4.test	28 801	4 244	2,93 %
VM5.test	28 801	4 289	2,93 %

TABLEAU 6.18 : Analyse des achats sur les jeux de données de VM Matériaux

9. Le jeu de données est blanchi aléatoirement dans le SGBD à l'aide d'une procédure SQL.

Validation croisée et comparaison

Tout d'abord, intéressons-nous à la validation des résultats sur les cinq ensembles d'apprentissage et les cinq ensembles de validation. Nous réalisons une extraction des règles avec un support $minSup$ égale à 0,0044 et une confiance $minConf$ égale à 60 %. Nous générons par la suite un ensemble de cohortes. Les premiers résultats sont récapitulés dans le tableau 6.19.

	VM1.base	**VM2.base**	**VM3.base**	**VM4.base**	**VM5.base**
# de règles	5 323	4 913	5 124	4 923	5 392
# de cohortes	74	73	72	73	76

TABLEAU 6.19 : Application de *CAPRE* sur les données de VM Matériaux

Pour maintenir une certaine robustesse, les cohortes composées d'au moins 2 règles sont conservées (cf. figure 6.15). Afin d'évaluer les *Top-20 ROI* recommandations, nous utilisons les deux métriques précision et rappel [75]. Nous privilégions la mesure de précision afin de minimiser le nombre de faux positifs, c'est-à-dire les clients que nous détecterions actionnables mais qui réellement n'ont pas acheté le produit. Enfin, nous présentons dans le tableau 6.20 les *Top-20* recommandations triées par ROI décroissant. Nous insistons sur le fait que sur l'ensemble des cohortes triées par précision moyenne, les *Top-10* recommandations possèdent une précision variant entre **71,64** et **80,17** et les *Bottom-10* recommandations entre **60,08** et **63,39**.

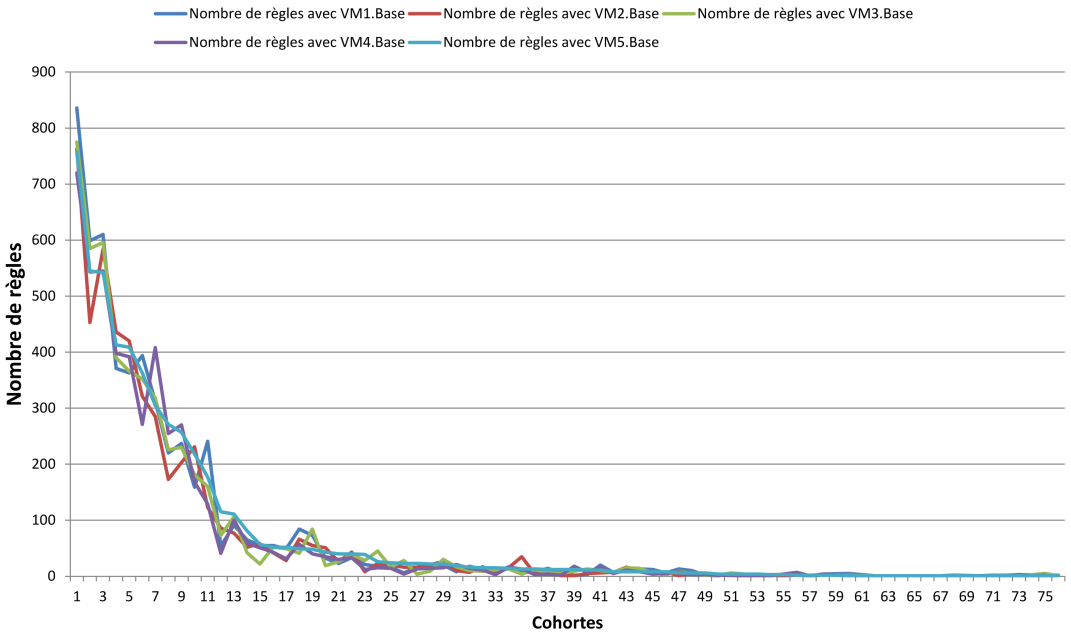


FIGURE 6.15 : Évolution du nombre de règles par cohorte sur *VM[1-5].base*

Top-20 ROI	# moyen de règles	$\overline{\text{Pr}}$	$\overline{\text{Rp}}$
Ciments gris	305	**68,51**	66,65
Blocs creux	363	**71,79**	78,01
Plaque de plâtre standard	757	**62,80**	92,83
Bétons courants standard	23	**71,60**	59,11
Menuiseries extérieures	65	**73,12**	47,67
Palettes facturées	413	**63,72**	69,79
Enduits mono-couches	4	**70,79**	17,39
Mortiers colles	111	**66,25**	65,08
Bétons hors normes	16	**66,64**	34,32
Rails et montants métalliques	545	**62,74**	89,69
Treillis bâtiment	20	**67,73**	36,79
Laine verre rouleaux	13	**70,77**	14,92
Tuiles romanes	2	**80,15**	28,52
Bois de charpente sapins	3	**69,38**	16,13
Laine verre panneaux	8	**70,77**	14,92
Plaque de plâtre hydrofuge	218	**54,66**	68,41
Colles et enduits pour plaques	542	**73,39**	76,34
Fourrures	363	**72,58**	88,48
Plâtres	14	**75,61**	19,47
Fers tort et ronds	12	**68,79**	20,43

TABLEAU 6.20 : Précision $\overline{\textbf{Pr}}$ et rappel $\overline{\textbf{Rp}}$ moyens des *Top-20 ROI* recommandations

Nous obtenons de bons résultats sur la précision moyenne des *Top-20 ROI* recommandations (cf. tableau 6.20). Par exemple la cohorte recommandant les blocs creux possède une précision moyenne de 71,79 % après validation croisée. De plus, les *Top-20 ROI* recommandations couvrent un large spectre de corps de métier :

- Les enduits (*Enduits mono-couches*) pour les enduiseurs ;
- Les tuiles (*Tuiles romanes*) pour les couvreurs ;
- Les plâtres (*Plâtres*) pour les plaquistes et plâtriers ;
- Les parpaings (*Blocs creux*) pour les maçons ;

Certaines cohortes permettent de détecter des anomalies dans la taxonomie produits de VM Matériaux (erreur de classification). Les experts métier proposent également une plus grande prise de risque sur certaines cohortes. En effet, l'ensemble de nos clients professionnels sont susceptibles de nous acheter des produits de quincaillerie. Dès lors, une fausse recommandation a un impact plus négligeable. En revanche, sur d'autres cohortes telle que les enduits mono-couches, les clients actionnables et profitables sont majoritairement des enduiseurs, un corps de métier spécifique, partageant peu de produits avec d'autres métiers. Une mauvaise recommandation d'enduits chez des électriciens ou plombiers par exemple pourrait avoir un effet néfaste sur le système.

6.2.7.2 Évaluation à l'aide d'un expert métier

Pour évaluer les résultats à l'aide d'un expert métier, nous avons créé un cube multidimensionnel permettant aisément à l'expert d'expliquer les recommandations et d'analyser les clients actionnables et profitables.

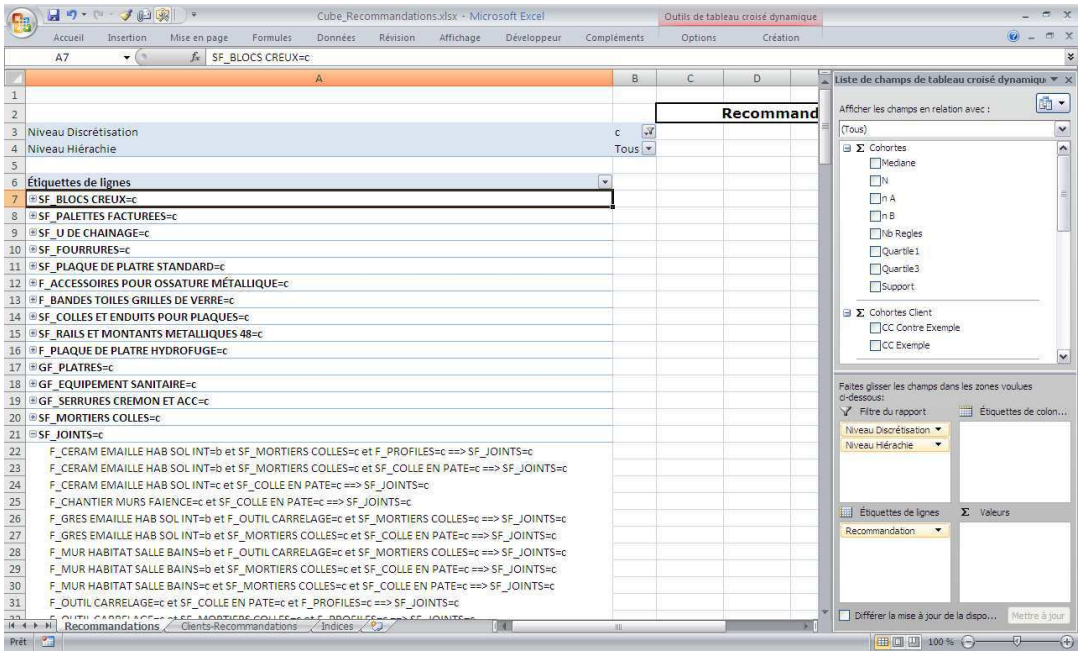


FIGURE 6.16 : Aperçu du cube *OLAP* des recommandations

Nous avons validé avec l'expert métier un certain nombre de recommandations actionnables en ciblant principalement les clients profitables. L'expert métier s'est appuyé sur l'ERP [10] et sur l'ensemble des rapports statiques et dynamiques qu'il a à sa disposition pour justifier ses avis sur les recommandations (cf. tableau 6.21).

40 clients actionnables et profitables ont ainsi été validés avec l'aide de l'expert métier (cf. tableau 6.21). Aucune recommandation inadaptée au client n'a été détectée. De nombreuses recommandations confortent les connaissances de l'expert (*connue*). Quelques recommandations sont qualifiées d'*inattendue* car l'expert métier n'y a pas pensé spontanément mais la connaissance s'avère pertinente. Par exemple, les *multi-spécialistes*, des clients professionnels qui se sont spécialisés lors d'une conjoncture difficile. Également, les clients qualifiés d'*eco-artisans* représentent des nouveaux métiers souvent difficiles à recommander pour les commerciaux.

10. *Enterprise Resource Planning*.

Identifiant client	Recommandation	Appréciation	Commentaire
PR152349	SF_BLOCS_CREUX=b	Connue	« Il s'agit bien d'un maçon qui s'approvisionne en blocs creux en fonction de la proximité entre le chantier et le magasin »
PR175278	SF_FOURRURES=c F_PLAQUE PLATRE HYDROFUGE=c SF_VISSERIE POINTES PLAQUES=c	Connue	« Il s'agit d'une entreprise de plaquistes de trois salariés très regardant sur les prix »
PR178605	GF_EQUIPEMENT SANITAIRE=c	Inattendue	« Il s'agit d'un chauffagiste qui ne travaille plus beaucoup avec VM Matériaux depuis un an »
PR290932	SF_CIMENTS GRIS=c SF_TREILLIS BATIMENT=c SF_CHAINAGES=c F_COUVERCLES BETON=c	Inattendue	« Il s'agit d'un maçon qui réalise un peu de travaux publics en fonction de la conjoncture »
PR033802	SF_BLOCS CREUX=b F_COUVERCLES BETON=c SF_TRACAGE=c	Inattendue	« Il s'agit d'un ancien paysagiste qui s'est spécialisé dans les travaux publics et qui achète l'ensemble de ses dallages et pavés extérieurs chez VM »
PR049247	SF_ACCESSOIRES TUBES PVC=c F_COUVERCLES BETON=c F_REGARDS CARRES=c F_REHAUSSES REGARDS=c	Connue	« Il s'agit bien de cinq ventes complémentaires. Généralement, les accessoires des tubes ne sont pas proposés en ventes croisées en magasin »
PR178572	SF_MORTIERS COLLES=c SF_JOINTS=c	Connue	« Il s'agit bien d'un carreleur qui n'achète que du carrelage chez VM Matériaux »
PR032579	SF_PLAQUE PLATRE STANDARD=c SF_BLOCS CREUX=b	Inattendue	« Il s'agit d'un maçon qui réalise de la rénovation en fonction de la saison et de la conjoncture »
PR034361	F_AUTRES TUILES TERRE CUITE=c	Inattendue	« Il s'agit bien d'un couvreur qui achète l'ensemble de ses tuiles chez VM Matériaux. Cependant, l'achat d'autres tuiles plus spécifiques s'avère plus rare »
PR281374	SF_RAILS ET MONTANTS=a SF_PLAQUE PLATRE STANDARD=a	Inattendue	« De plus en plus, les menuisiers plaquent pour améliorer leur marge et gagner du temps sur les chantiers »

TABLEAU 6.21 : Exemples de validation à l'aide d'un expert métier

6.3 Conclusion

Dans ce chapitre, nous avons présenté une application réelle de nos travaux sur l'entrepôt de données du groupe VM Matériaux. Nous avons réalisé un ciblage des clients appétents à participer à une campagne commerciale affinant ainsi la base client sur laquelle la méthodologie *CAPRE* est appliquée. Nous avons appliqué notre méthodologie à l'aide de l'outil *ARKIS* sur environ 10 000 clients et 100 000 produits. La phase d'actionnabilité souligne l'intérêt de réaliser des recommandations pour la gestion de la relation client. La phase de profitabilité est une approche optimiste qui permet de trier les recommandations et les clients par profit. La validation statistique des résultats et l'aide d'un expert métier souligne la qualité des recommandations et la minimisation des fausses recommandations.

7 Conclusion et perspectives

SOMMAIRE

7.1 Bilan

Dans un contexte concurrentiel, les entreprises cherchent à trouver des gisements de rentabilité. Comme il est plus rentable de fidéliser un client existant que d'en acquérir un nouveau, les objectifs de résultats se transforment en objectifs de ventes sur les clients existants. De ce fait, les entreprises doivent tenir compte en priorité des exigences de leurs clients pour les fidéliser. C'est à cet objectif que souhaite répondre la Gestion de la Relation Client. Avec l'avènement des nouvelles technologies de l'information et de la communication, les sources de données sont devenues très nombreuses et très riches. Pour comprendre le comportement de leurs clients, les entreprises mettent en œuvre des outils et des techniques pour extraire des connaissances sur les clients : c'est l'Extraction de Connaissances à partir des Données. Ces connaissances et leur actionnabilité fournissent aux experts métier un outil d'aide à la décision dont la performance peut être mesurée par le ROI généré par les actions. Les systèmes de recommandation sont une solution adaptée pour mettre en place ces outils car ils permettent de filtrer l'information puis de recommander de manière proactive des produits susceptibles de fidéliser le client.
Les travaux présentés ici s'inscrivent dans le cadre d'une stratégie commerciale de fidélisation au travers des forces de vente. Dans ce cadre, les recommandations qui sont utilisées sont dites « intrusives », une mauvaise recommandation pouvant en effet avoir des répercussions importantes sur le client. De plus, le commercial peut refuser d'utiliser le système s'il ne juge pas les recommandations suffisamment pertinentes. Dès lors, démarcher et recommander de manière profitable pour développer la valeur client s'avère être une tâche difficile dans la pratique. C'est pour s'affranchir de ces contraintes que nous avons proposé la méthodologie *CAPRE* de recommandations actionnables et profitables. Les contributions de nos travaux sont résumées comme suit :

- Nous avons développé une nouvelle **méthodologie** qui s'appuie sur un modèle à base de cohortes de règles qui permettent d'**expliquer** et d'**interpréter** les recommandations ;

- Nous avons incorporé un **modèle économique** d'actionnabilité/profitabilité des recommandations fondé sur la similarité entre les exemples et les contre-exemples ainsi que sur des critères métier ;

- Nous avons **développé** un prototype de recherche *ARKIS* implémentant notre méthodologie ;

- Nous avons testé et **validé** l'efficacité de notre méthodologie sur le jeu de données de référence *MovieLens* ;

- Nous avons **appliqué** notre prototype de recherche et notre méthodologie sur les données opérationnelles du groupe VM Matériaux.

Les spécificités de la méthodologie *CAPRE*

Nous avons proposé une méthodologie pour les systèmes de recommandation fondée sur l'analyse des chiffres d'affaires des clients sur des familles de produits, nommée *CAPRE* (*Customer Actionability and Profitability Recommendation*). Cette méthodologie consiste à extraire des comportements de référence sous la forme de cohortes de règles d'association et à en évaluer l'actionnabilité et l'intérêt économique. Les recommandations sont réalisées en ciblant les contre-exemples les plus actionnables sur les règles les plus rentables. Les spécificités de notre approche sont les suivantes :

- *CAPRE* permet de pointer les clients présentant un manque à gagner pour lesquels faire une recommandation est objectivement raisonnable (*actionnabilité*) ;
- *CAPRE* permet de quantifier le manque à gagner (*profitabilité*) et ainsi d'allouer des moyens en proportion pour réaliser la recommandation (*canaux de communication*) ;
- *CAPRE* fournit des modèles de recommandation explicites (*boîte blanche*) : les forces de vente disposent de règles pour comprendre l'origine de la recommandation et peuvent analyser la population de clients exemples dont les comportements cohérents ont amené à l'apparition de la cohorte dans les données.

Les notions d'*actionnabilité* et de *profitabilité* constituent une originalité forte de notre méthodologie. Deux stratégies s'offrent aux experts : démarcher les clients actionnables et/ou suggérer les recommandations profitables. Dans les deux cas, fidéliser sur le long terme passe par une succession de recommandations à court terme.

Le modèle économique d'actionnabilité/profitabilité

Avec leur capacité d'achat potentiellement inexploité, les contre-exemples des règles peuvent contribuer théoriquement au développement du CA des entreprises. Cependant, tous les clients ne présentent pas forcément la même réceptivité face à une recommandation. Notre méthodologie se concentre sur les clients les plus actionnables, c'est-à-dire les contre-exemples les plus « proches » des exemples vis-à-vis de leur comportement d'achat et de leurs variables descriptives. Plus un contre-exemple est proche des exemples, plus il est probable qu'il se comporte comme un exemple, c'est-à-dire qu'il développe davantage son CA pour un produit. Pour être actionnable, un client (i) ne doit pas présenter un parcours d'achat « extrême » sur les prémisses des règles de la recommandation par rapport aux exemples et (ii) être suffisamment proche de certains exemples vis-à-vis de ses variables descriptives.
Pour les clients pour lesquels faire une recommandation est objectivement raisonnable, un manque à gagner en Euros peut être quantifié pour la recommandation appliquée. Deux approches optimistes de calculs de la profitabilité ont été proposées, estimant la somme qui aurait due être dépensée si le contre-exemple s'était comporté comme un exemple. Ces approches ont l'avantage de trier les recommandations et les clients par profit et ainsi d'aider les experts métier à prioriser les produits à recommander et les clients à démarcher.

L'outil de recommandations *ARKIS*

Nous avons développé un outil opérationnel implémentant toutes les étapes décrites dans la méthodologie *CAPRE*. Cet outil nommé *ARKIS* (*Association Rule Knowledge Interactive Search*) permet d'extraire des recommandations actionnables et profitables pour les experts métier. *ARKIS* a été développé en Java et est interopérable avec un SGBD. L'interface graphique mise en place permet aux *data miners* et aux experts métier d'extraire des règles de comportements d'achats et de générer des recommandations sous forme de cohortes actionnables et profitables. Les tests d'utilisation d'*ARKIS* sur des données réelles montrent que l'outil aide à découvrir des recommandations et des clients actionnables et profitables.

La validation sur les données de référence *MovieLens* et les données opérationnelles de VM Matériaux

Une partie importante de notre travail a été de valider l'efficacité de notre méthodologie. Pour y parvenir, nous avons utilisé le jeu de données de référence *MovieLens*. En exploitant les ensembles d'apprentissage, nous avons généré des prédictions de recommandations et comparé les résultats avec les valeurs réelles contenues dans les ensembles de validation. Les résultats des *Top-10* recommandations obtenus présentent une bonne précision. Les résultats sur les *Bottom-10* sont appréciables. L'agrégation des règles en cohortes permet de construire un modèle plus robuste qu'une simple extraction de règles d'association, réduisant ainsi le nombre de recommandations et le nombre de fausses de recommandations.

Nous avons présenté le retour d'expérience du projet de fouille de données mené chez VM Matériaux pour améliorer le retour sur investissement d'opérations commerciales. Nous avons réalisé un ciblage de clients susceptibles de participer à une opération commerciale. L'estimation du profit a été pertinente. Malgré la conjoncture qui s'est matérialisée par une baisse du CA sur les clients habituellement routés, la méthode a permis de sauvegarder le CA de l'opération commerciale. Au cours de la campagne ciblée, 115 nouveaux clients présents dans notre liste de routage ont participé, représentant un chiffre d'affaires additionnel de 1 200 000 €.

Nous avons appliqué notre méthodologie *CAPRE* sur les clients précédemment ciblés et plus de 100 000 références produits. 23 578 règles d'association ont été extraites, générant 174 cohortes. Une validation croisée sur les données blanchies aléatoirement a permis d'obtenir de bons résultats, présentant une précision appréciable sur les *Top-20 ROI* recommandations. L'avis d'un expert métier du Négoce de matériaux a permis d'évaluer les recommandations. Les cohortes validées se sont avérées être des connaissances connues ou inattendues par l'expert et aucune fausse recommandation n'a été détectée. Nous avons quantifié les actions de recommandations sur quelques cohortes sélectionnées par les experts. Le profit espéré laisse prévoir un réel potentiel de développement de la valeur client, répondant ainsi à notre objectif premier, fidéliser.

7.2 Perspectives

Approfondir les mesures d'actionnabilité et de profitabilité

Nous envisageons d'améliorer la mesure d'actionnabilité en incorporant des pondérations dans la mesure de distance. Le fait que certaines variables aient plus de poids que d'autres dans la mesure des proximités entre clients est en effet conforme à l'intuition des experts. De manière générale, le critère d'actionnabilité peut être modulé par la *prise de risque* que souhaite prendre l'entreprise. Une entreprise présentant des clients « non risqués » (souvent le cas des sites de commerce en ligne) sera moins exigeante sur l'élagage des contre-exemples. En revanche, une entreprise respectant la loi de *Pareto* [160] illustrant que généralement 20 % des clients représentent environ 80 % de l'activité de l'entreprise, aura tendance à être prudente sur une partie de son portefeuille client.
Nous estimons également pertinent d'améliorer la mesure de profitabilité en considérant le nombre de contre-exemples actionnables de chaque cohorte. En effet, il peut être plus difficile de démarcher 1 000 clients 1 000 € profitables que 100 clients 10 000 € profitables. Les coûts afférents ne sont pas forcément proportionnels. De plus, la marge nette des produits et les valeurs de stocks courants pourraient influer sur le système de recommandation, réconfortant ainsi les experts et les objectifs métier de l'entreprise. Enfin, la profitabilité de chaque contre-exemple pourrait être pondérée par la qualité des règles dont il est issu.

Segmenter pour recommander

De nombreuses entreprises possèdent non seulement une typologie produits mais également une typologie de clients. Nous songeons à réaliser une segmentation préalable des clients pour appliquer notre méthodologie *CAPRE* sur des groupes de clients plus homogènes, et ainsi pouvoir utiliser des catégories de produits plus précises. Le client se verra ainsi recommander des produits plus spécifiques en fonction de son segment d'appartenance. Néanmoins, cette méthode pourrait restreindre la transversalité de l'approche et pourrait masquer des comportements atypiques que nous aurions détecté initialement.

Rétro-action et rétro-profit

Les retours d'expérience des experts et notamment l'explication métier des fausses recommandations constituent des éléments pertinents pour rétroagir sur les recommandations. Par exemple, le moteur de recommandation *Genius* d'*Apple* permet à l'utilisateur de supprimer ses recommandations d'applications directement sur son mobile. Dès lors, le système impacte sa base de recommandations.
De la même manière, le retour sur investissement de recommandations peut ne pas être rentable pour l'entreprise bien que l'estimation du profit ait été profitable. Nous envisageons de pondérer les recommandations non seulement en fonction du profit estimé mais également du profit effectif. Ainsi, le système rétroagit en fonction des

gains réalisés par l'entreprise. Nous pourrions imaginer que le retour sur investissement de certaines recommandations aide au déclenchement de recommandations plus onéreuses.
Dans tous les cas, il faut que le système conserve aussi son rôle d'aiguillon pour les commerciaux, en continuant de fournir des recommandations même lorsque celles-ci sont optimistes (exemple d'un client qui achèterait systématiquement un certain produit à la concurrence). Il y a donc un certain équilibre à trouver entre rétro-action et déviation par rapport au marché réel.

Et du côté applicatif...

Nos perspectives applicatives consistent à poursuivre nos travaux dans le domaine de la fouille de données appliquée aux systèmes de recommandation.
À court terme, notre souhait est d'améliorer l'outil *ARKIS* en rendant interopérable l'ensemble du processus avec tout système d'information d'entreprise. Nous souhaitons également interconnecter l'outil avec un ERP pour utiliser les règles de recommandations et aider à la mise en place d'opérations commerciales ciblées par l'intermédiaire des forces de vente commerciales.
À moyen terme, nous cherchons à optimiser le système pour recommander en temps réel. L'idée est d'actionner les recommandations dès la détection d'un client dans nos magasins et d'assister les forces commerciales pour appuyer leur argumentaire de vente.
À plus long terme, nous souhaitons constituer une base de données des recommandations *success stories* sur lesquelles un retour sur investissement effectif a été mesuré. Nous enrichissons ainsi nos bases de données de connaissances clients / produits issues des recommandations. Dès lors, les collaborateurs du groupe VM Matériaux pourront consulter les connaissances qui ont effectivement été profitables.
Ceci constitue bien évidemment une liste non exhaustive des perspectives applicatives pour VM Matériaux. Notre souhait étant de traiter en priorité une intégration et un déploiement d'*ARKIS* dans le système d'information et une adhésion forte des utilisateurs. Nous avons la conviction que ce point constitue l'élément essentiel d'un système de recommandation fiable et performant.

Annexes

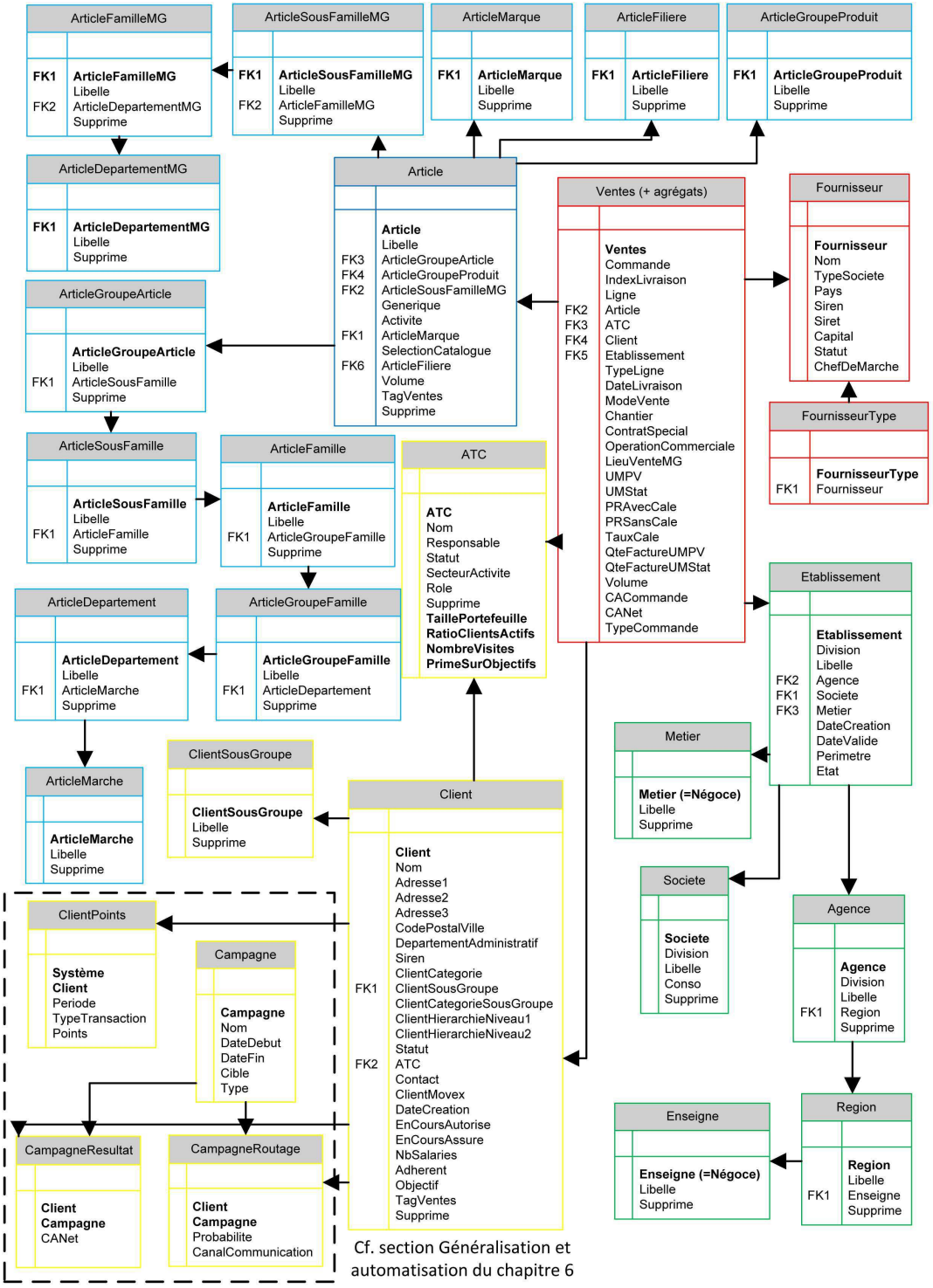
A

Vue simplifiée de l'entrepôt de données de VM Matériaux

Cette annexe présente une vue simplifiée de la partie gestion commerciale de l'entrepôt de données de VM Matériaux. L'entrepôt est composé approximativement de 180 tables pour une volumétrie de 93 gigaoctets. Nous listons dans le tableau A.1 quelques caractéristiques des cinq tables les plus volumineuses de la gestion commerciale.

Nom de la table	**Nombre de lignes**	**Nombre de colonnes**	**Place occupée (*KB*)**
Ventes	28 302 535	30	10 131 632
Client	1 110 283	51	564 600
Article	422 776	21	81 096
Fournisseur	32 920	23	5 000
ATC	1 826	15	216

TABLEAU A.1 : Volumétrie des tables les plus représentatives de la gestion commerciale de l'entrepôt de données de VM Matériaux


ArticleFamilleMG
FK1 ArticleFamilleMG
FK2 ArticleDepartementMG
Libelle
Supprime
ArticleSousFamilleMG
FK1 ArticleSousFamilleMG
FK2 ArticleFamilleMG
Libelle
Supprime
ArticleMarque
FK1 ArticleMarque
Libelle
Supprime
ArticleFiliere
FK1 ArticleFiliere
Libelle
Supprime
ArticleGroupeProduit
FK1 ArticleGroupeProduit
Libelle
Supprime
ArticleDepartementMG
FK1 ArticleDepartementMG
Libelle
Supprime
Article
Article
Libelle
FK3 ArticleGroupeArticle
FK4 ArticleGroupeProduit
FK2 ArticleSousFamilleMG
Generique
Activite
FK1 ArticleMarque
SelectionCatalogue
FK6 ArticleFiliere
Volume
TagVentes
Supprime
Ventes (+ agrégats)
Ventes
Commande
IndexLivraison
Ligne
FK2 Article
FK3 ATC
FK4 Client
FK5 Etablissement
TypeLigne
DateLivraison
ModeVente
Chantier
ContratSpecial
OperationCommerciale
LieuVenteMG
UMPV
UMStat
PRAvecCale
PRSansCale
TauxCale
QteFactureUMPV
QteFactureUMStat
Volume
CACommande
CANet
TypeCommande
Fournisseur
Fournisseur
Nom
TypeSociete
Pays
Siren
Siret
Capital
Statut
ChefDeMarche
FournisseurType
FK1 FournisseurType
Fournisseur
ArticleGroupeArticle
FK1 ArticleGroupeArticle
Libelle
ArticleSousFamille
Supprime
ArticleSousFamille
FK1 ArticleSousFamille
Libelle
ArticleFamille
Supprime
ArticleFamille
FK1 ArticleFamille
Libelle
ArticleGroupeFamille
Supprime
ATC
ATC
Nom
Responsable
Statut
SecteurActivite
Role
Supprime
TaillePortefeuille
RatioClientsActifs
NombreVisites
PrimeSurObjectifs
ArticleDepartement
FK1 ArticleDepartement
Libelle
ArticleMarche
Supprime
ArticleGroupeFamille
FK1 ArticleGroupeFamille
Libelle
ArticleDepartement
Supprime
Etablissement
Etablissement
Division
Libelle
FK2 Agence
FK1 Societe
FK3 Metier
DateCreation
DateValide
Perimetre
Etat
Metier
Metier (=Négoce)
Libelle
Supprime
ArticleMarche
ArticleMarche
Libelle
Supprime
ClientSousGroupe
ClientSousGroupe
Libelle
Supprime
Client
Client
Nom
Adresse1
Adresse2
Adresse3
CodePostalVille
DepartementAdministratif
Siren
ClientCategorie
FK1 ClientSousGroupe
ClientCategorieSousGroupe
ClientHierarchieNiveau1
ClientHierarchieNiveau2
Statut
FK2 ATC
Contact
ClientMovex
DateCreation
EnCoursAutorise
EnCoursAssure
NbSalaries
Adherent
Objectif
TagVentes
Supprime
Societe
Societe
Division
Libelle
Conso
Supprime
Agence
Agence
Division
Libelle
FK1 Region
Supprime
ClientPoints
Système
Client
Periode
TypeTransaction
Points
Campagne
Campagne
Nom
DateDebut
DateFin
Cible
Type
Enseigne
Enseigne (=Négoce)
Libelle
Supprime
Region
Region
Libelle
FK1 Enseigne
Supprime
CampagneResultat
Client
Campagne
CANet
CampagneRoutage
Client
Campagne
Probabilite
CanalCommunication
Cf. section Généralisation et automatisation du chapitre 6

B

Rappels sur la théorie statistique de Vladimir VAPNIK

Cette annexe présente une synthèse des travaux de la théorie de Vapnik [285] à la base des fondements de l'outil *KXEN* [94, 224].

B.1 Notations

Nous disposons de données d'apprentissage $(x^1, y^1), (x^2, y^2), ..., (x^n, y^n)$ où le vecteur $x^i = (x^i_1, x^i_2, ..., x^i_p)$ est une observation et y^i est la variable cible à expliquer. y peut être une variable discrète (classification) ou continue (régression). Les (x^i, y^i) sont supposés être un échantillon de tirages i.i.d (*Independent and Identically-Distributed*) issus d'une distribution fixe mais inconnue $P(X, Y)$.

Pour expliquer la variable cible, on utilise une classe de fonctions dépendant d'un paramètre θ : $\Phi_\theta = \{f(., W, \theta), W \in \aleph\}$. Il s'agit par exemple de la classe des polynômes de degré θ, $\aleph$ étant l'espace des coefficients du polynôme et W un vecteur contenant ces coefficients. Le paramètre θ permet de sélectionner des familles de fonctions plus ou moins complexes alors que W est le paramètre que l'on utilise lors de l'apprentissage pour un θ fixé. Un modèle issu de cette classe produit pour chaque observation x une sortie $y = f(x, W, \theta)$. À partir des données $(x^1, y^1), (x^2, y^2), ..., (x^n, y^n)$, nous cherchons le *meilleur* modèle $\hat{y} = f(x, \hat{W}, \theta)$ produit par un certain algorithme ou principe d'inférence et correspondant au meilleur paramètre $\hat{W}$.

B.2 Coût et risque

Notons tout d'abord une fonction de perte $L[y, f(x, W, \theta)]$ mesurant le coût qu'il y a, à remplacer la vraie valeur y par la valeur calculée $f(x, W, \theta)$. L'erreur d'apprentissage ou *risque empirique* est alors défini comme le coût moyen sur l'ensemble d'apprentissage :

$$R_{emp}(W, \theta) = \frac{1}{n} \sum_{i=1}^{n} L[y^i, f(x^i, W, \theta)] \tag{B.1}$$

L'*erreur en généralisation* est le *coût moyen théorique* sur l'ensemble de la population, c'est-à-dire l'erreur attendue sur de nouvelles données :

$$R_{Gen}(W, \theta) = \int L[y, f(x, W, \theta)].dP(x, y) \tag{B.2}$$

On utilise classiquement comme coût l'*écart quadratique* :

$$L[y, f(x, W, \theta)] = [y - f(x, W, \theta)]^2 \tag{B.3}$$

Dans le cas du coût de l'écart quadratique, le *risque empirique* est l'écart quadratique moyen MSE (*Mean Square Error*) : $R_{emp}(W, \theta) = \frac{1}{n} \sum_{i=1}^{n} [y^i - f(x^i, W, \theta)]^2$.

B.3 Minimisation du risque empirique

Le principe d'inférence est la minimisation du risque empirique (*Empirical Risk Minimization* ou ERM). L'utilisation du principe ERM permet de déterminer le meilleur $\hat{W}$ (*data fit*) mais pas de choisir θ.
À partir d'un ensemble d'apprentissage $(x^1, y^1), (x^2, y^2), ..., (x^n, y^n)$, le principe d'inférence ERM consiste à minimiser le risque empirique, c'est-à-dire à maximiser la précision sur l'ensemble d'apprentissage. La figure B.1 expose plusieurs modèles pour les observations. Le principe ERM nous amène à choisir le troisième qui a la meilleure précision.

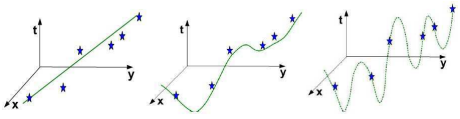


FIGURE B.1 : Précision [94]

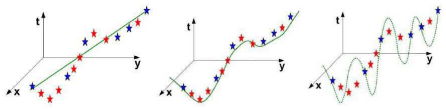


FIGURE B.2 : Robustesse [94]

La figure B.2 révèle le comportement du modèle $f(x, W, \theta)$ sur de nouvelles données, un ensemble de test par exemple. Pour privilégier la robustesse, c'est-à-dire la qualité du modèle sur de nouvelles données, le deuxième modèle devrait être choisi. Cette illustration montre que le principe ERM seul ne peut pas garantir précision et robustesse.

B.4 Dimension de VAPNIK CHERVONENKIS

La dimension de VAPNIK CHERVONENKIS ou VC dimension, mesure la capacité de modélisation de la classe de fonctions Φ_θ. Dans un souci de simplicité, je présenterai ce concept dans le cas d'une classification en deux classes. Soit, un échantillon de n observations $(x^1, x^2, ..., x^n)$ en p variables : $x^i = (x_1^i, x_2^i, ..., x_p^i)$. Il y a 2^n façons de séparer ces n observations en deux classes. On dit que la famille de fonctions $\Phi_\theta = \{f(., W, \theta), W \in \aleph\}$ *pulvérise* l'échantillon si toutes les 2^n séparations sont réalisables (avec un $\hat{W}_\theta$ bien choisi). La famille Φ_θ est de VC dimension h_θ si h_θ est le nombre maximum de points qui peut être pulvérisé par Φ_θ :

- Il existe au moins un échantillon de h_θ observations qui peut être pulvérisé par Φ_θ.
- Aucun échantillon de $h_\theta + 1$ observations ne peut être pulvérisé par Φ_θ.

Par exemple, si on utilise la famille des droites de $\mathbf{R}^2$, la figure B.3 montre que la VC dimension de cette famille est 3 :

- Il y a au moins un échantillon de 3 points qui peut être pulvérisé par les droites.
- Aucun échantillon de 4 points ne peut être pulvérisé par les droites.

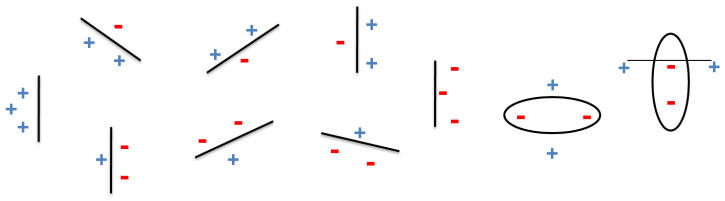


FIGURE B.3 : Illustration de la VC dimension dans $\mathbf{R}^2$

B.5 *Statistical Learning Theory*

La *Statistical Learning Theory* de Vladimir Vapnik [285] est une théorie générale qui utilise la VC dimension et repose sur quatre principes :

- **Robustesse** : la capacité à généraliser correctement sur de nouvelles données. On dit que le principe d'inférence ERM est robuste pour la classe de fonctions $\Phi_\theta = \{f(., W, \theta), W \in \aleph\}$ si et seulement si $R_{emp}(\theta)$ et $R_{Gen}(\theta)$ convergent vers la même limite quand la taille de l'échantillon n tend vers l'infini. C'est le cas si la famille Φ_θ est de VC dimension h finie [285].

- **Vitesse de convergence** : capacité à généraliser de mieux en mieux quand le nombre de données d'apprentissage augmente. Vladimir Vapnik [285] a démontré que quel que soit $\eta \in [0,1]$, alors, avec une probabilité $1-\eta$,

$$R_{Gen}(\theta) \leq R_{emp}(\theta) + \epsilon(n,h) \; avec, \; \epsilon(n,h) = \sqrt{\frac{1+\ln(\frac{2n}{h})}{\frac{n}{h}} - \frac{\ln\eta}{n})} \qquad \text{(B.4)}$$

Ce résultat est indépendant de la distribution $P(X,Y)$ de (X,Y) : il démontre que, si n est assez grand, $\epsilon \simeq 0$ et donc l'erreur en généralisation est du même ordre que l'erreur d'apprentissage, c'est-à-dire que le modèle est robuste (cf. figure B.4).

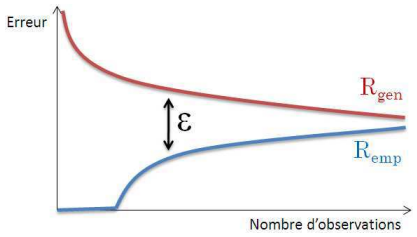


FIGURE B.4 : Robustesse de l'ERM

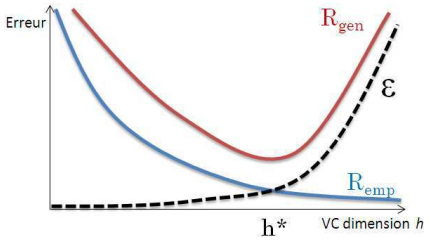


FIGURE B.5 : Capacité de généralisation

- **Contrôle de la capacité de généralisation** : stratégie qui permet de contrôler la capacité de généralisation à partir des seules données d'apprentissage. En pratique, on ne peut pas rendre n (la taille de la base dont on dispose) aussi grand que nécessaire. On voit donc dans l'équation B.4 ci-dessus que deux alternatives s'offrent à nous :
  - Quand $\frac{n}{h}$ est grand, on minimise le risque empirique R_{emp} et on est assuré que R_{Gen} est du même ordre.
  - Quand $\frac{n}{h}$ est petit, on doit minimiser les deux termes : R_{emp} et $\epsilon(n,h)$. La figure B.5 montre que, à n fixé, quand h augmente, l'écart ϵ tend vers l'infini et donc, à partir d'une dimension h^* l'erreur en généralisation R_{Gen} se met à croître et devient de plus en plus différente de R_{emp}. Le point h^* du minimum de R_{Gen} correspond au meilleur compromis entre précision (R_{emp} petit) et robustesse (R_{Gen} petit).

- **Stratégie pour obtenir de bons algorithmes** : nous venons de voir qu'il existe une valeur optimale h^* qui réalise le meilleur compromis entre précision et robustesse : il faut une stratégie qui permette de l'obtenir. Vladimir Vapnik [285] introduit pour cela la SRM (*Structural Risk Minimization*) utilisant des familles de fonctions emboitées $\Phi_{\theta_1} \subset \Phi_{\theta_2} \subset ... \subset \Phi_{\theta_k} \subset ...$ de VC dimension croissante $h_1 < h_2 < ... < h_k < ....$ L'algorithme pour déterminer le modèle est le suivant : on découpe l'ensemble de données en deux parties, l'une est dite ensemble d'apprentissage et l'autre ensemble de validation. Parfois, on découpe en trois

parties, avec un *ensemble de test*, uniquement pour mesurer les performances du modèle produit. L'erreur en validation comme estimateur de l'erreur de généralisation est utilisée.

1. Commencer avec Φ_{θ_1}.
2. *Fit* des données : pour chaque $\Phi\theta_k$, faire :
   - Sur l'ensemble d'apprentissage, produire le meilleur modèle de $\Phi\theta_k$, c'est-à-dire choisir : $\hat{W_{\theta_k}} = arg \min_W R_{emp}(W, \theta_k)$
   - Mesurer l'erreur sur l'ensemble de validation

$$R_{Val}(\hat{W}_{\theta_k}) = \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} L[y^i, f(x^i, \hat{W}_{\theta_k}, \theta_k)] \tag{B.5}$$

   - Si $R_{val}(\hat{W}_{\theta_{k-1}}) \gg R_{val}(\hat{W_{\theta_k}})$ alors faire $k = k + 1$ et aller à 2, sinon stop et faire $\theta_{k^*} = \theta_k$
3. Choix du modèle : le meilleur modèle est celui qui correspond à θ_{k^*}.

La *Statistical Learning Theory* apporte un ensemble de résultats permettant de contrôler la famille de modèles comprenant la solution, par le biais de la VC dimension h de la famille retenue. Cette méthode de contrôle garantit le meilleur compromis précision/robustesse du modèle obtenu. Les résultats étant indépendants de la distribution des données, on s'affranchit de la nécessité de connaître cette distribution et de l'estimer. La SRM ne donne aucune indication sur la « bonne » classe de modèles, sauf que la VC dimension doit être finie.

B.6 Mise en œuvre de la *SRM* dans KXEN

Une validation croisée est réalisée aussi bien au cours de l'encodage que de la modélisation. Le jeu de données est divisé en trois sous ensembles (cf. figure B.6) pour l'apprentissage, la validation (choix du modèle) et le test pour mesurer les performances du modèle final.

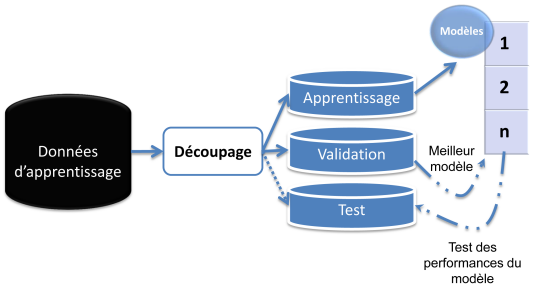


FIGURE B.6 : Échantillonnage du jeu de données

C

Évaluation des modèles dans KXEN

Cette annexe présente la méthodologie mise en place pour évaluer les modèles de *scoring* au sein de l'outil KXEN [223].

C.1 *Scoring* d'appétence client

Pour prédire la cible binaire de participation à une campagne marketing, nous choisissons de calculer un score à l'aide de la technique de la régression *ridge*.

Régression *ridge*

La régression *ridge* [189] présente comme avantages de pénaliser les paramètres lorsque la variable est fortement bruitée et d'être peu sensible aux corrélations. Lorsque les variables prédictives sont corrélées, la régression utilise cette corrélation pour compenser les effets de chaque variable. La régression *ridge* peut fournir la contribution des variables, i.e. des poids polynomiaux W_x, soulignant la contribution relative C_x de chaque variable x dans le modèle [274] :

$$C_x = W_x / \sum_x W_x \tag{C.1}$$

Dans nos travaux, nous considérons que la variable cible binaire désigne la participation à une opération commerciale, c'est-à-dire à l'achat d'un produit (1 pour acheter, 0 sinon). L'objectif est de prévoir les acheteurs d'une opération commerciale dans une population de clients [27]. Étant donné un seuil s, on prédit qu'un client i est un acheteur si le score s_i calculé par le modèle est supérieur à s (cf. figure C.1). On note $u(s)$, la proportion de clients dont le score calculé par le modèle est supérieur à s :

$$u(s) = P(s_i \geq s) \tag{C.2}$$

On note $v(s)$ la proportion d'acheteurs réels détectés par le modèle :

$$v(s) = P(s_i \geq s \mid i = \text{acheteur}) \tag{C.3}$$

		Réel	
		Acheteur	Non Acheteur
Prédit	Acheteur (score $\geq s$)	Vrai Positif	Faux Positif
	Non Acheteur (score $< s$)	Faux Négatif	Vrai Négatif

TABLEAU C.1 : Matrice de confusion

Le modèle permet d'obtenir une fonction s_i de scores décroissants pour chaque client i reflétant la probabilité p_i (obtenue par normalisation de s_i) d'acheter durant la campagne marketing.

Précision et robustesse

Afin que les experts métier puissent visualiser la précision et la robustesse de nos modèles, nous utilisons des courbes *lift* (courbes C_3 et C_4 sur la figure C.1).
Une courbe *lift* (variante de la courbe ROC) est une courbe paramétrique qui représente la proportion d'acheteurs détectés $v(s)$ en fonction de la proportion de clients sélectionnés $u(s)$ [108]. Elle est construite en triant les clients par ordre de score décroissant. Par ailleurs, il peut être profitable à l'expert métier de visualiser la « représentation du *lift* » présentant cette fois-ci le taux d'augmentation du *lift* en ordonnée, la courbe étant par conséquent décroissante.
La précision et la robustesse d'un modèle peuvent être mesurées en comparant la courbe *lift* à une courbe aléatoire et une courbe idéale (courbes C_2 et C_1 sur la figure C.1). La courbe aléatoire est la courbe $y = x$ (on détecte α % des acheteurs en sélectionnant α % des clients). La courbe idéale est celle dans laquelle tous les acheteurs sont sélectionnés en premier.
À partir de la courbe *lift*, deux indicateurs peuvent être calculés. Le premier indicateur est équivalent à l'indice de GINI [85], nommé KI dans KXEN. Il correspond à l'aire entre la courbe de validation et la courbe aléatoire. Le KI de validation est égal au rapport des aires $C/(A+B+C)$ et le KI d'apprentissage est égale à $(B+C)/(A+B+C)$. Il mesure la précision du modèle, c'est-à-dire la capacité des variables d'entrée à expliquer la cible. L'indicateur compris entre 0 (modèle purement aléatoire) et 1 (modèle parfait) permet de classer les modèles en fonction de leur pouvoir explicatif face à la variable à expliquer. En effet, l'aire totale sous la courbe *lift* (AUL) est corrélée à l'aire totale sous la courbe ROC (AUC) par la formule suivante : $AUL = f/2 + (1-f)(AUC)$ où f est la fréquence a priori de l'événement dans l'ensemble de la population. Cela signifie que :

$$KI = \frac{AUL - \frac{1}{2}}{\frac{1-f}{2}} = \frac{f + 2(1-f)AUC - 1}{1-f} = 2AUC - 1 \tag{C.4}$$

Le deuxième indicateur, nommé KR dans l'outil KXEN, correspond à la différence d'aire entre les deux courbes de *lift* d'apprentissage et de validation, soit $(1-B)/(A+B+C)$. Il mesure la robustesse du modèle, c'est-à-dire sa capacité à fournir le même niveau de qualité sur un nouveau jeu de données, typiquement le jeu de données de validation. Il est également compris entre 0 et 1 et il est préférable qu'il soit supérieur à 0,95 pour que le modèle soit robuste.
Par exemple, sur la figure C.1, le point M montre que sur l'ensemble d'apprentissage, en ciblant 50 % des clients (les meilleurs selon le modèle), on détecte 80 % des acheteurs.

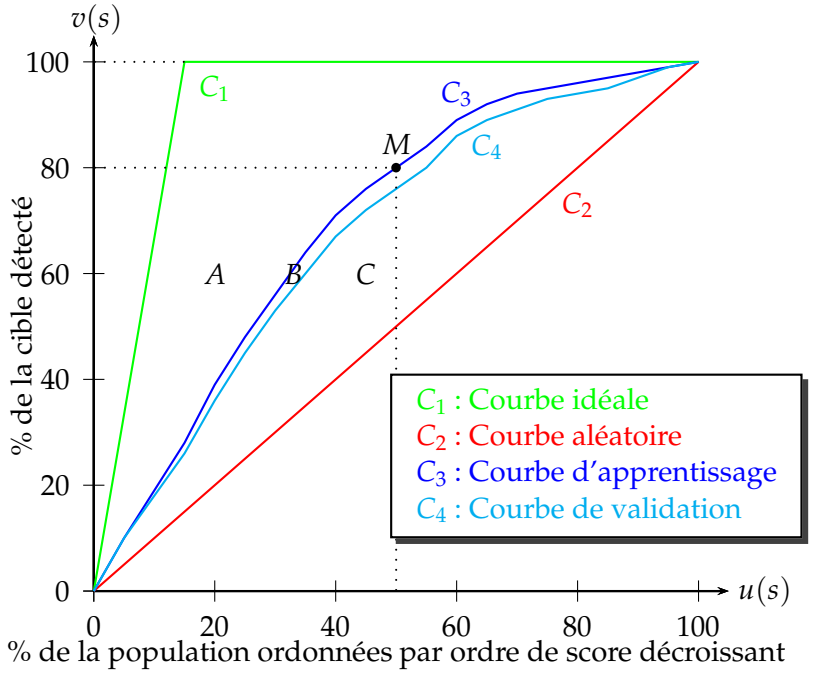


FIGURE C.1 : Courbe *lift*

Cette partie de la méthodologie permet d'obtenir deux indicateurs KI et KR synthétiques représentant la courbe *lift*. Dans la section suivante, la courbe de profit naïve permet d'introduire une contrainte économique dans le ciblage client.

Courbe de profit naïve

De manière à ce que les experts métier puissent estimer le retour sur investissement engendré par un modèle sur une opération commerciale, nous utilisons des courbes de profit naïves sur l'ensemble d'apprentissage ou de validation. Une courbe de profit est la transformation d'une courbe de *lift* à l'aide d'une matrice des coûts (cf. figure C.2) définie par les experts métier.
Le profit naïf pour une campagne marketing peut être défini de la manière suivante : il s'agit de la marge nette réalisée en contactant une proportion $u(s)$ % de clients. Soit N le nombre de clients dans l'échantillon étudié, G la marge nette moyenne générée

		Réel	
		Acheteur	Non Acheteur
Prédit	Acheteur (score $\geq s$)	$G - H$	$0 - H$
	Non Acheteur (score $< s$)	0	0

TABLEAU C.2 : Matrice des coûts

par client et H la dépense moyenne de communication par client (cf. tableau C.2) :

$$\begin{aligned} ProfitNaif(s) = N * [&P(i = acheteur | s(i) \geq s) * (G - H) \\ &- P(i = non\ acheteur | s(i) \geq s) * H] \end{aligned} \tag{C.5}$$

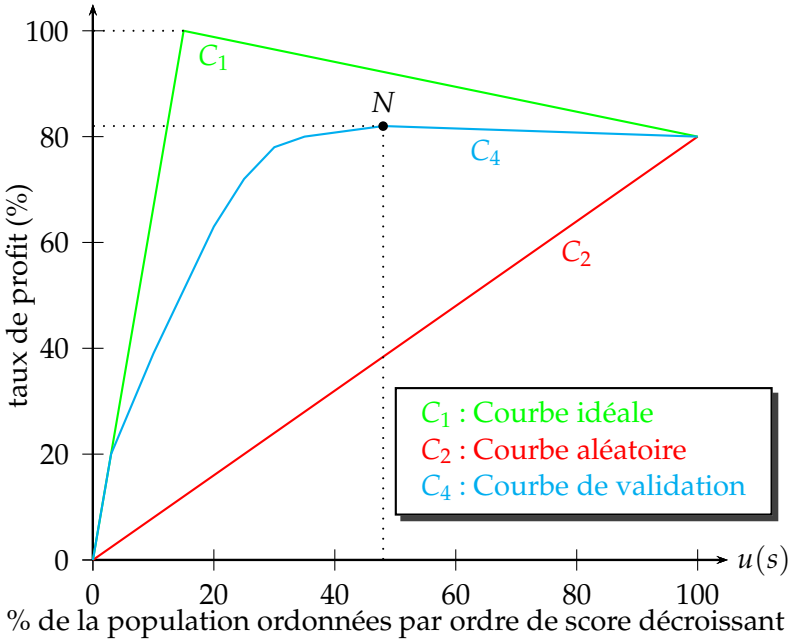


FIGURE C.2 : Courbe de profit naïve

Le profit maximal théorique, $profitMAX$, est le modèle où tous les acheteurs sont sélectionnés en premier. Ainsi, une courbe de profit naïve (cf. figure C.2) est une courbe paramétrique qui représente le taux de profit ($ProfitNaif(s)/profitMAX$) en fonction de la proportion de clients sélectionnés $u(s)$. Cette courbe présente une ordonnée différente de la courbe de *lift* avec non plus le pourcentage d'acheteurs détectés mais le pourcentage de ROI maximal de manière à mesurer graphiquement le retour financier de l'opération commerciale. Par exemple, le point N sur la figure C.2 signifie que sur l'ensemble de validation, il faut contacter 48 % de la population pour obtenir un ROI maximal égal à 82 % du profit maximal théorique. Cette partie de

notre méthodologie permet ainsi d'obtenir le point optimal (ordonnée maximale) sur la courbe indiquant la proportion de la population qui doit être contactée.

C.2 Estimation de la marge nette par client

Pour prédire la marge nette g_i par client i, nous choisissons d'utiliser un second modèle dont la cible est continue.

Changement de cible métier

Généralement, le score généré par la régression *ridge* peut être très ambigu pour une interprétation *directe* par les décideurs métier. Cette restriction est résolue par l'utilisation d'un seconde résultat : l'estimation de la marge nette générée par client. Nous considérons une variable cible continue représentant la somme de marge nette g_i réalisée par un client i durant une campagne marketing. L'objectif de ce second modèle est d'estimer la marge nette g_i par visite client durant la campagne. L'ensemble d'apprentissage est moins volumineux que dans le cas de la régression avec une cible binaire puisqu'il ne correspond qu'aux clients ayant participé à la campagne sur laquelle a lieu l'apprentissage du modèle. Les clients importants à forte marge nette représentent une faible partie de l'ensemble d'apprentissage. Il s'avère donc difficile de déterminer leur comportement d'achat. Le modèle généré est ainsi moins précis et constitue une option secondaire pour trier les listes de routage. Un modèle avec une cible continue ne peut pas être analysé grâce à une courbe *lift* ou à une courbe de profit naïve (cf. figures C.1 et C.2).

Fusion des listes de routage

Nous pouvons combiner les résultats des modèles (binaire et continu) pour optimiser le routage des clients. Les scores générés par la première régression sont discrétisés (10^{-4} de précision) et permettent de classer les clients par ordre décroissant. Ensuite, nous complétons cette classification avec le résultat de la seconde régression *ridge*, i.e. la marge nette. Pour les clients présentant des scores très proches issus de la première régression, l'ordre des clients est modifié en fonction de la marge nette estimée. Le directeur marketing et le directeur achats de VM Matériaux ont décidé d'attacher plus d'importance à un nouveau client (i.e. à la valeur p_i) qu'au développement de la marge nette par client (i.e. à la valeur g_i).

Bibliographie

[1] N. M. Adams, D. J. Hand, and R. J. Till. Mining for Classes and Patterns in Behavioural Data. *The Journal of the Operational Research Society, Special Issue : Credit Scoring and Data Mining*, 52(9) :1017–1024, September 2001.

[2] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems : A survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions On Knowledge and Data Engineering*, 17(6) :734–749, 2005.

[3] C. C. Aggarwal, J. L. Wolf, K.-L. Wu, and P. S. Yu. Horting Hatches an Egg : A New Graph-Theoretic Approach to Collaborative Filtering. In *In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge discovery and data mining*, pages 201–212. ACM Press, 1999.

[4] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington D.C., May 1993.

[5] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on VLDB*, pages 487–499, Santiago, Chile, September 1994.

[6] S. R. Ahmed. Applications of Data Mining in Retail Business. In *ITCC '04 : Proceedings of the International Conference on Information Technology : Coding and Computing (ITCC'04) Volume 2*, page 455, Washington, DC, USA, 2004. IEEE Computer Society.

[7] C. Anderson. *The Long Tail : Why the Future of Business Is Selling Less of More*. Hyperion, 2006.

[8] E. Anderson and V. Mittal. Strengthening the Satisfaction-Profit Chain. *Journal of Service Research*, 3(2) :107–120, 2000.

[9] J. L. Anderson, L. D. Jolly, and A. E. Fairhurst. Customer Relationship Management in Retailing : A Content Analysis of Retail Trade Journals. *Journal of Retailing and Consumer Services*, 14(6) :394 – 399, 2007. Data Mining Applications in Retailing and Consumer Services.

[10] C. Apté. Data Mining : An Industrial Research Perspective. *Computing in Science and Engineering*, 4(2) :6–9, 1997.

[11] C. Argyris. *Knowledge for Action : A Guide to Overcoming Barriers to Organizational Change*. Jossey-Bass Inc., 1993.

[12] C. Argyris and D. A. Schon. *Organizational Learning : A Theory of Action Perspective (Addison-Wesley Series on Organization Development)*. Addison-Wesley, June 1978.

[13] M. Z. Ashrafi, D. Taniar, and K. Smith. Redundant Association Rules Reduction Techniques. *Int. J. Bus. Intell. Data Min.*, 2(1) :29–63, 2007.

[14] M.-J. Avenier. L'Élaboration de Savoirs Actionnables en PME Légitimés Dans une Conception des Sciences de Gestion Comme des Sciences de l'Artificiel. *Revue internationale P.M.E*, 17(3-4) :13–42, 2004.

[15] B. Baesens, S. Viaene, and J. Vanthienen. Post-Processing of Association Rules. Open access publications from katholieke universiteit leuven, Katholieke Universiteit Leuven, 2000.

[16] M. Balabanovic and Y. Shoham. Fab : Content-Based, Collaborative Recommendation. *Communication of the ACM*, 40(3) :66–72, 1997.

[17] P. Barwise and J. U. Farley. The State of Interactive Marketing in Seven Countries : Interactive Marketing Comes of Age. *Journal of Interactive Marketing*, 19(3) :67–80, 2005.

[18] Y. Bastide, N. Pasquier, R. Taouil, S. Gerd, and L. Lakhal. Mining Minimal Non-redundant Association Rules Using Frequent Closed Itemsets. In *Proceedings of the First International Conference on Computational Logic*, CL '00, pages 972–986, London, UK, 2000. Springer-Verlag.

[19] Y. Bastide, R. Taouil, N. Pasquier, S. Gerd, and L. Lakhal. Mining Frequent Patterns With Counting Inference. *SIGKDD Explor. Newsl.*, 2 :66–75, December 2000.

[20] R. J. Bayardo. Efficiently Mining Long Patterns from Databases. In *SIGMOD Conference*, pages 85–93, 1998.

[21] R. J. Bayardo and R. Agrawal. Mining the Most Interesting Rules. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pages 145–154. ACM Press, 1999.

[22] A. L. C. Bazzan. Agents and Data Mining in Bioinformatics : Joining Data Gathering and Automatic Annotation with Classification and Distributed Clustering. In *ADMI*, pages 3–20, 2009.

[23] M. Bécue-Bertaut and J. Pagès. Multiple Factor Analysis and Clustering of a Mixture of Quantitative, Categorical and Frequency Data. *Computational Statistics & Data Analysis*, 52(6) :3255–3268, 2008.

[24] R. Bell, Y. Koren, and C. Volinsky. Modeling Relationships at Multiple Scales to Improve Accuracy of Large Recommender Systems. In *KDD '07 : Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 95–104, New York, NY, USA, 2007. ACM.

[25] R. M. Bell and Y. Koren. Lessons From the Netflix Prize Challenge. *SIGKDD Explor. Newsl.*, 9 :75–79, December 2007.

[26] F. Bentayeb, C. Favre, and O. Boussaid. A User-Driven Data Warehouse Evolution Approach for Concurrent Personalized Analysis Needs. *Integr. Comput.-Aided Eng.*, 15(1) :21–36, 2008.

[27] M. J. Berry and G. Linoff. *Data Mining : Techniques Appliquées au Marketing, à la Vente et aux Services Clients*. Masson, 1997.

[28] A. Berson, S. Smith, and K. Thearling. *Building Data Mining Applications for CRM*. McGraw-Hill Professional, 1999.

[29] I. Bhattacharya, S. Godbole, A. Gupta, A. Verma, J. Achtermann, and K. English. Enabling Analysts in Managed Services for CRM Analytics. In *KDD '09 : Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1077–1086, New York, NY, USA, 2009. ACM.

[30] M. Bilgic. Explaining Recommendations : Satisfaction vs. Promotion. In *In Proceedings of Beyond Personalization 2005, the Workshop on the Next Stage of Recommender Systems Research(IUI2005*, pages 13–18, 2005.

[31] G. Bitran and S. Mondschein. Mailing Decisions in the Catalog Sales Industry. *Management Science*, 42 :1364–1381, 1996.

[32] J. Blanchard. *Un Système de Visualisation Pour l'Extraction, l'Évaluation, et l'Exploration interactives des Règles d'Association*. Thèse, Sciences et Technologies de l'Information et des Matériaux, Nantes, November 2005. Henri Briand (Dir.).

[33] J. Blanchard, F. Guillet, H. Briand, and R. Gras. Assessing Rule Interestingness With a Probabilistic Measure of Deviation from Equilibrium. In *Proceedings of the 11th international symposium on Applied Stochastic Models and Data Analysis ASMDA-2005*, pages 191–200. ENST, 2005.

[34] Z. Bodie, A. Kane, and A. Marcus. *Essentials of Investments*. McGraw-Hill/Irwin, 6 edition, November 2005.

[35] C. Bothorel and M. Bouklit. Détection de Structures de Communauté Dans les Hyper-réseaux d'Interactions. In David Simplot-Ryl and Sebastien Tixeuil, editors, *10ème Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications (AlgoTel'08)*, pages 57–60, Saint-Malo France, 2008.

[36] R. J. Brachman and T. Anand. The Process of Knowledge Discovery in Databases. pages 37–57, 1996.

[37] J. S. Breese, D. Heckerman, and C. Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. pages 43–52. Morgan Kaufmann, 1998.

[38] T. Breur. How to evaluate campaign response The relative contribution of data mining models and marketing execution. *Journal of Targeting, Measurement and Analysis for Marketing*, 15 :103–112(10), March 2007.

[39] H. Briand, M. Sebag, R. Gras, and F. Guillet. *Mesures de qualité pour la fouille de données*. Cépaduès, 2004.

[40] S. Brin, R. Motwani, and C. Silverstein. Beyond Market Baskets : Generalizing Association Rules to Correlations. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, SIGMOD'97, pages 265–276, New York, NY, USA, 1997. ACM.

[41] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In *SIGMOD '97 : Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 255–264, New York, NY, USA, 1997. ACM.

[42] S. Brown. *Customer Relationship Management, La Gestion de La Relation Client*. PriceWaterHouseCoopers, 2006.

[43] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He. Music Recommendation by Unified Hypergraph : Combining Social Media Information and Music Content. In *Proceedings of the international conference on Multimedia*, MM '10, pages 391–400, New York, NY, USA, 2010. ACM.

[44] B. G. Buchanan and E. H. Shortliffe. *Rule Based Expert Systems : The Mycin Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1985.

[45] J. R. Bult and T. Wansbeek. Optimal Selection for Direct Mail. *Marketing Science*, 14 :378–394, 1995.

[46] D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, and T. Yiu. MAFIA : A Maximal Frequent Itemset Algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 17 :1490–1504, 2005.

[47] R. Burke. Knowledge-Based Recommender Systems. In *Encyclopedia of Library and Information Systems*, volume 69, 2000.

[48] R. Burke. Hybrid Recommender Systems : Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12 :331–370, November 2002.

[49] R. Burke. The Adaptive Web. chapter Hybrid Web Recommender Systems, pages 377–408. Springer-Verlag, Berlin, Heidelberg, 2007.

[50] C. H. Cai, A. W. C. Fu, C. H. Cheng, and W. W. Kwong. Mining Association Rules with Weighted Items. In *Proceedings of the 1998 International Symposium on Database Engineering and Applications*, page 68, Washington, DC, USA, 1998. IEEE Computer Society.

[51] L. Candillier, F. Meyer, and M. Boullé. Comparing State-of-the-Art Collaborative Filtering Systems. In *Proceedings of the 5th international conference on Machine Learning and Data Mining in Pattern Recognition*, MLDM '07, pages 548–562, Berlin, Heidelberg, 2007. Springer-Verlag.

[52] L. Cao. Domain-Driven, Actionable Knowledge Discovery. *IEEE Intelligent Systems*, 22(4) :78–89, 2007.

[53] L. Cao. Domain Driven Data Mining (D3M). In *ICDMW 08 : Proceedings of the 2008 IEEE International Conference on Data Mining Workshops*, pages 74–76, Washington, DC, USA, 2008. IEEE Computer Society.

[54] L. Cao. Domain-Driven Data Mining : Challenges and Prospects. *IEEE Transactions on Knowledge and Data Engineering*, 22 :755–769, 2010.

[55] L. Cao, C. Luo, and C. Zhang. Developing Actionable Trading Agents. *Intelligent Agent Technology, IAT '07 IEEE/WIC/ACM International Conference on*, pages 72–75, November 2007.

[56] L. Cao, D. Luo, and C. Zhang. Knowledge Actionability : Satisfying Technical and Business Interestingness. *Int. J. Bus. Intell. Data Min.*, 2(4) :496–514, 2007.

[57] L. Cao and Y. Ou. Market Microstructure Patterns Powering Trading and Surveillance Agents. *Journal of Universal Computer Science*, 14(14) :2288–2308, 2008.

[58] L. Cao, P. Yu, C. Zhang, and H. Zhang. *Data Mining for Business Applications*. Springer, 2009.

[59] L. Cao and C. Zhang. Domain-Driven Actionable Knowledge Discovery in the Real World. In *PAKDD*, pages 821–830, 2006.

[60] L. Cao and C. Zhang. Domain-Driven Data Mining : A Practical Methodology. In *International Journal of Data Warehousing and Mining (IJDDWM)*, volume 2, pages 49–65. IGI Global, 2006.

[61] L. Cao and C. Zhang. The Evolution of KDD : Towards Domain-Driven Data Mining. *International Journal of Pattern Recognition and Artificial Intelligence*, 21(4) :677–692, 2007.

[62] L. Cao, C. Zhang, Y. Zhao, P. S. Yu, and G. Williams. DDDM2007 : Domain Driven Data Mining. *SIGKDD Explor. Newsl.*, 9(2) :84–86, 2007.

[63] L. Cao, Y. Zhao, H. Zhang, D. Luo, C. Zhang, and E. Park. Flexible Frameworks for Actionable Knowledge Discovery. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints), 2009.

[64] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization : Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[65] S. Castagnos. *Modélisation de Comportements et Apprentissage Stochastique non Supervisé de Stratégies d'Interactions Sociales au Sein de Systèmes Temps Réels de Recherche et d'Accès à l'Information*. Thèse, IAEM-Lorraine, LORIA Laboratory (KIWI Team), Nancy, France, November 2008. Anne Boyer (Dir).

[66] S. Castagnos and A. Boyer. A Client/Server User-Based Collaborative Filtering Algorithm : Model and Implementation. In *Proceeding of the 2006 conference on ECAI 2006 : 17th European Conference on Artificial Intelligence August 29 – September 1, 2006, Riva del Garda, Italy*, pages 617–621, Amsterdam, The Netherlands, The Netherlands, 2006. IOS Press.

[67] S. Castagnos, N. Jones, and P. Pu. Eye-Tracking Product Recommenders' Usage. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 29–36, New York, NY, USA, 2010. ACM.

[68] A. Ceglar and J. F. Roddick. Association Mining. *ACM Comput. Surv.*, 38, July 2006.

[69] S. H. S. Chee, J. Han, and K. Wang. RecTree : An Efficient Collaborative Filtering Method. In *Proceedings of the Third International Conference on Data Warehousing and Knowledge Discovery*, DaWaK '01, pages 141–151, London, UK, 2001. Springer-Verlag.

[70] M.-S. Chen, J. Hun, P. S. Yu, I. T. J, and W. R. Ctr. Data mining : An Overview From a Database Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8 :866–883, 1996.

[71] Y. Chen, D. Pavlov, P. Berkhin, A. Seetharaman, and A. Meltzer. Practical Lessons of Data Mining at Yahoo ! In *CIKM '09 : Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1047–1056, New York, NY, USA, 2009. ACM.

[72] D. Chickering, D. Geiger, and D. Heckerman. Learning Bayesian Networks : Search Methods and Experimental Results. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 112–128, 1995.

[73] Y. H. Cho, J. K. Kim, and S. H. Kim. A Personalized Recommender System Based on Web Usage Mining and Decision Tree Induction. *Expert Systems with Applications*, 23(3) :329 – 342, 2002.

[74] R. L. Cilibrasi and P. M. B. Vitanyi. The Google Similarity Distance. *IEEE Trans. on Knowl. and Data Eng.*, 19 :370–383, March 2007.

[75] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is Seeing Believing ? : How Recommender System Interfaces Affect Users' Opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '03, pages 585–592, New York, NY, USA, 2003. ACM.

[76] M. Cosquer, F. Gros, and A. Livartowski. Qualité et Datawarehouse dans le Milieu Hospitalier. In *Revue Nationale des Technologies de l'Information (RNTI)*, volume E1, page 107. Cépaduès, 2004.

[77] C. Cumby, A. Fano, R. Ghani, and M. Krema. Predicting Customer Shopping Lists from Point-of-Sale Purchase Data. In *Proceedings of the tenth ACM*

SIGKDD international conference on Knowledge discovery and data mining, KDD '04, pages 402–409, New York, NY, USA, 2004. ACM.

[78] M. Czarkowski. A Scrutable Adaptive Hypertext. In *Proceedings of the Fourth Workshop on Empirical Evaluation of Adaptive Systems, held at the 10th International Conference on User Modeling UM2005*, pages 384–387. Springer, 2005.

[79] P. Danaher and J. Rossiter. A Comparison of the Effectiveness of Marketing Communication Channels : Perspectives from Both Receivers and Senders. *Autralian and New Zealand Marketing Academy (ANZMAC)*, July 2006.

[80] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google News Personalization : Scalable Online Collaborative Filtering. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 271–280, New York, NY, USA, 2007. ACM.

[81] T. H. Davenport and J. G. Harris. *Competing on Analytics : The New Science of Winning*. Harvard Business School Press, Boston, MA, USA, 2007.

[82] A. David, A. Hatchuel, and R. Laufer. *Les Nouvelles Fondations des Sciences de Gestion*. November 2008.

[83] C. Derquenne, S. Goutier, S. Lembo, and V. Stéphan. Tirer Profit des Sources Externes pour l'Enrichissement des Bases Clients - Application du Data Mining Prédictif aux Bases EDF. In *RNTI*, 2003.

[84] M. Deshpande and G. Karypis. Item Based Top-N Recommendation Algorithms. *ACM Transactions on Information Systems*, 22 :143–177, 2004.

[85] C. Drummond and R. Holte. Cost Curves : An Improved Method for Visualizing Classifier Performance. In *Machine Learning*, volume 65, pages 95–130. Springer Netherlands, October 2006.

[86] B. Escofier. Traitement Simultané de Variables Qualitatives et Quantitatives en Analyse Factorielle. *Les cahiers de l'analyse des données*, 4(2) :137–146, 1979.

[87] K. J. Ezawa and T. Schuermann. Fraud/Uncollectible Debt Detection Using a Bayesian Network Based Learning System : A Rare Binary Outcome with Mixed Data Structures. In P. Besnard and S. Hanks, editors, *UAI*, pages 157–166. Morgan Kaufmann, 1995.

[88] U. Fayyad. A Data Miner's Story - Getting to Know the Grand Challenges. Invited Talk, KDD 2007.

[89] U. Fayyad, G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann, 2001.

[90] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.

[91] R. Feraud, M. Boullé, F. Clérot, and F. Fessant. Vers l'Exploitation de Grandes Masses de Données. In *Revue Nationale des Technologies de l'Information (RNTI)*, pages 241–252. Cépaduès, 2008.

[92] A. Fernandez. *Les Nouveaux Tableaux de Bord des Managers*. 4 edition, 2008.

[93] S. M. Finlay. Towards Profitability : a Utility Approach to the Credit Scoring Problem. *Journal of the Operational Research Society*, 59 :921–931(11), July 2008.

[94] F. Fogelman-Soulié and E. Marcadé. Industrial Mining of Massive Data Sets. *NATO Science for Peace and Security Series, Information and Communication Security*, 19 :44–61, 2008.

[95] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge Discovery in Databases : an Overview. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 1–30. AAAI/MIT Press, 1991.

[96] A. A. Freitas. On Rule Interestingness Measures. *Knowledge-Based Systems*, 12(5-6) :309 – 315, 1999.

[97] R. Garber. An Interview with Ronald A. Howard. *Decision Analysis*, 6(4) :263–272, 2009.

[98] L. Geng and H. J. Hamilton. Interestingness Measures for Data Mining : A Survey. *ACM Comput. Surv.*, 38(3) :9, 2006.

[99] C. Giraud-Carrier and O. Povel. Characterising Data Mining Software. *Intell. Data Anal.*, 7(3) :181–192, 2003.

[100] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using Collaborative Filtering to Weave an Information Tapestry. *Commun. ACM*, 35(12) :61–70, 1992.

[101] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste : A Constant Time Collaborative Filtering Algorithm. *Inf. Retr.*, 4 :133–151, July 2001.

[102] N. Good, J. B. Schafer, J. A. Konstan, A. Borchers, B. Sarwar, J. Herlocker, and J. Riedl. Combining Collaborative Filtering with Personal Agents for Better Recommendations. In *In Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 439–446, 1999.

[103] O. Grabova, J. Darmont, J.-H. Chauchat, and I. Zolotaryova. Business Intelligence for Small and Middle-Sized Entreprises. *SIGMOD Record*, 39(2) :39–50, June 2010.

[104] W. Graco, T. Semenova, and E. Dussobarsky. Toward Knowledge-Driven Data Mining. In *ACM SIGKDD Workshop on Domain Driven Data Mining*, pages 49–54, 2007.

[105] G. Grahne and J. Zhu. Fast Algorithms for Frequent Itemset Mining Using FP-Trees. *IEEE Trans. on Knowl. and Data Eng.*, 17 :1347–1362, October 2005.

[106] R. Gras, E. Suzuki, F. Guillet, and Filippo. *Statistical Implicative Analysis*. Studies in Computational Intelligence 127, 2008, Springer, 2008.

[107] S. Greco, B. Matarazzo, N. Pappalardo, and R. Slowinski. Measuring Expected Effects of Interventions Based on Decision Rules. *J. Exp. Theor. Artif. Intell.*, 17(1-2) :103–118, 2005.

[108] Q. Gu, L. Zhu, and Z. Cai. Study on Measure Criteria in Evaluating Classification Performance : Lift charts, ROC and Precision-Recall Curves. In S. Zeng, Y. Liu, Q. Zhang, and L. Kang, editors, *Progress In Intelligence Computation and Applications*, pages 488–492. China Univ Geosciences Press, 2007.

[109] F. Guillet and H. J. Hamilton. *Quality Measures in Data Mining*. Springer, 2007.

[110] C. L. Gunnarsson, M. M. Walker, V. Walatka, and K. Swann. Lessons Learned : A Case Study Using Data Mining in the Newspaper Industry. *The Journal of Database Marketing*, 14(38) :271–280(10), July 2007.

[111] S. Gupta, D. R. Lehmann, and J. A. Stuart. Valuing Customers. *Journal of Marketing Research*, 41(1)(HBS Marketing Research Paper No. 03-08) :7–18, February 2004.

[112] S. H. Ha, S. M. Bae, and S. C. Park. Customer's Time-Variant Purchase Behavior and Corresponding Marketing Strategies : an Online Retailer's Case. *Computers and Industrial Engineering*, 43(4) :801 – 820, 2002.

[113] J. Hahn, R. Kauffman, and J. Park. Designing for ROI : Toward a Value-Driven Discipline for E-commerce Systems Design. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 7 - Volume 7*, HICSS '02, page 200, Washington, DC, USA, 2002. IEEE Computer Society.

[114] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent Pattern Mining : Current Status and Future Directions. *Data Mining and Knowledge Discovery*, 14(1), 2007.

[115] J. Han, H. Gonzalez, X. Li, and D. Klabjan. Warehousing and Mining Massive RFID Data Sets. In X. Li, O. Zaïane, and Z.-h. Li, editors, *Advanced Data Mining and Applications*, volume 4093 of *Lecture Notes in Computer Science*, pages 1–18. Springer Berlin / Heidelberg, 2006.

[116] J. Han and M. Kamber. *Data mining : Concepts and Techniques*. Morgan Kaufmann, 2 edition, 2000.

[117] J. Han, J. Wang, Y. Lu, and P. Tzvetkov. Mining Top-K Frequent Closed Patterns without Minimum Support. In *In Proceedings of ICDM 02*, pages 211–218, 2002.

[118] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, MA, USA, 2001.

[119] P. E. Hart and J. Graham. Query-Free Information Retrieval. *IEEE Expert : Intelligent Systems and Their Applications*, 12 :32–37, September 1997.

[120] Z. He, X. Xu, J. Z. Huang, and S. Deng. Mining Class Outliers : Concepts, Algorithms and Applications in CRM. *Expert Systems with Applications*, 27(4) :681 – 697, 2004.

[121] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency Networks for Inference, Collaborative Filtering, and Data Visualization. *J. Mach. Learn. Res.*, 1 :49–75, September 2001.

[122] V. Henning and J. Reichelt. Mendeley - A Last.fm For Research ? In *Proceedings of the 2008 Fourth IEEE International Conference on eScience*, pages 327–328, Washington, DC, USA, 2008. IEEE Computer Society.

[123] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An Algorithmic Framework for Performing Collaborative Filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 230–237, New York, NY, USA, 1999. ACM.

[124] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining Collaborative Filtering Recommendations. pages 241–250, 2000.

[125] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. Inf. Syst.*, 22 :5–53, January 2004.

[126] G. Herschel. *CRM Analytics Scenario : The Emergence of Integrated Insight*. Gartner Customer Relationship Management Summit, 2006.

[127] T. Highley and P. Reynolds. Marginal Cost-Benefit Analysis for Predictive File Prefetching. In *Proc. of the 41st Annual ACM Southeast Conference*, 2003.

[128] W. Hill, L. Stead, M. Rosenstein, and G. Furnas. Recommending and Evaluating Choices in a Virtual Community of Use. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '95, pages 194–201, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[129] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for Association Rule Mining – A General Survey and Comparison. *SIGKDD Explorations*, 2(2) :1–58, 2000.

[130] E. Horvitz and M. Barry. Display of Information for Time-Critical Decision Making. In *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 296–305. Morgan Kaufmann, 1995.

[131] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse. The Lumière Project : Bayesian User Modeling for Inferring the Goals and Needs of Software Users. In *In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 256–265. Morgan Kaufmann, 1998.

[132] M. Hosseini-Pozveh, M. A. Nematbakhsh, and N. Movahhedinia. A Multidimensional Approach for Context-Aware Recommendation in Mobile Commerce. *CoRR*, 2009. informal publication.

[133] C.-H. Hsu. Data mining to Improve Industrial Standards and Enhance Production and Marketing : An Empirical Study in Apparel Industry. *Expert Syst. Appl.*, 36(3) :4185–4191, 2009.

[134] C.-N. Hsu, H.-H. Chung, and H.-S. Huang. Mining Skewed and Sparse Transaction Data for Personalized Shopping Recommendation. *Mach. Learn.*, 57 :35–59, October 2004.

[135] W.-C. Hu. Adaptive Web Browsing Using Web Mining Technologies for Internet Enabled Mobile Handheld Devices. *Emerging Trends and Challenges in Information Technology Management*, page 4, 2006.

[136] J. Huang and C. X. Ling. Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17 :299–310, 2005.

[137] F. Hussain, H. Liu, E. Suzuki, and H. Lu. Exception Rule Mining with a Relative Interestingness Measure. In *PADKK '00 : Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, pages 86–97, London, UK, 2000. Springer-Verlag.

[138] T. Imielinski and A. Virmani. Association Rules... and What's Next ? Towards Second Generation Data Mining Systems. In *ADBIS '98 : Proceedings of the Second East European Symposium on Advances in Databases and Information Systems*, pages 6–25, London, UK, 1998. Springer-Verlag.

[139] M. Jamali and M. Ester. TrustWalker : a Random Walk Model for Combining Trust-Based and Item-Based Recommendation. In *KDD '09 : Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 397–406, New York, NY, USA, 2009. ACM.

[140] M. Jambu. *Introduction au Data Mining, Analyse Intelligente des Données*. Eyrolles, 1999.

[141] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender Systems An Introduction*. Cambridge University Press, 2011.

[142] Y. Jiang, K. Wang, A. Tuzhilin, and A. Fu. Mining Patterns that Respond to Actions. In *Fifth IEEE International Conference on Data Mining*, pages 669–672. IEEE Computer Soc, 2005.

[143] R.-S. Kaplan and D.-P. Norton. *Le tableau de Bord Prospectif*. 2 edition, 2003.

[144] D. A. Keim. Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1) :1–8, 2002.

[145] D. A. Keim, F. Mansmann, J. Schneidewind, H. Ziegler, and J. Thomas. Visual Analytics : Scope and Challenges. December 2008. Visual Data Mining : Theory, Techniques and Tools for Visual Analytics, Springer, Lecture Notes In Computer Science (lncs).

[146] S. Kelly. Mining Data to Discover Customer Segments. *Interactive Marketing*, 4 :235–242(8), March 2003.

[147] C. Kim and J. Kim. A Recommendation Algorithm Using Multi-Level Association Rules. In *Proceedings of the 2003 IEEE/WIC Int. Conf. on Web Intelligence*, WI '03, page 524, Washington, DC, USA, 2003. IEEE Computer Society.

[148] J. W. Kincaid. *Customer Relationship Management : Getting it Right*. Prentice Hall Press, Upper Saddle River, NJ, USA, 2003.

[149] J. Kleinberg. Challenges in Social Network Data : Processes, Privacy and Paradoxes, 2007.

[150] A. Knott, A. Hayes, and S. A. Neslin. Next-Product-to-Buy Model for Cross-Selling Applications. *Journal of Interactive Marketing*, 16 :59–75, 2002.

[151] Y. Kodratoff. L'Extraction de Connaissances à Partir de Données : Un Nouveau Sujet Pour la Recherche Scientifique. In *Revue Électronique sur l'Apprentissage par les Données*, volume 1, 1997.

[152] R. Kohavi. Focus the Mining Beacon : Lessons and Challenges from the World of E-Commerce. In *PKDD*, page 7, 2005.

[153] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. GroupLens : Applying Collaborative Filtering to Usenet News. *Commun. ACM*, 40 :77–87, March 1997.

[154] B. Krulwich and C. Burkey. The InfoFinder Agent : Learning User Interests Through Heuristic Phrase Extraction. *IEEE Expert*, 12(5) :22–27, 1997.

[155] A. Kusiak. Data mining : Manufacturing and Service Applications. *International Journal of Production Research*, 44 :4175–4191, 2006.

[156] R. D. Lawrence, G. S. Almasi, V. Kotlyar, M. S. Viveros, and S. S. Duri. Personalization of Supermarket Product Recommendations. *Data Min. Knowl. Discov.*, 5 :11–32, January 2001.

[157] R. Lefébure and G. Venturi. *Gestion de la Relation Client*. Eyrolles, 2005.

[158] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich. A Multicriteria Decision Aid for Interestingness Measure Selection. Technical report, LUSSI Department, GET/ENST, Bretagne, France, May 2004.

[159] P. Lenca, B. Vaillant, P. Meyer, and S. Lallich. Association Rule Interestingness Measures : Experimental and Theoretical Studies. In *Quality Measures in Data Mining*, pages 51–76. 2007.

[160] J. Lendrevie, J. Lévy, and D. Lindon. *MERCATOR : Théorie et Pratique du Marketing*. 9 edition, 2009.

[161] C. W.-k. Leung, S. C.-f. Chan, and F.-l. Chung. A Collaborative Filtering Framework Based on Fuzzy Association Rules and Multiple-Level Similarity. *Knowledge and Inf. Syst.*, 10 :357–381, 2006.

[162] C. W.-k. Leung, S. C.-f. Chan, and F.-l. Chung. An Empirical Study of a Cross-Level Association Rule Mining Approach to Cold-Start Recommendations. *Knowledge-Based Systems*, 21(7) :515 – 529, 2008.

[163] J. Li. On Optimal Rule Discovery. *IEEE Trans. on Knowl. and Data Eng.*, 18(4) :460–471, 2006.

[164] J. Li, B. Tang, and N. Cercone. Applying Association Rules for Interesting Recommendations Using Rule Templates. In *PAKDD*, pages 166–170, 2004.

[165] S. Li, B. Sun, and R. Wilcox. Cross-Selling Sequentially Ordered Products : An Application to Consumer Banking Services. *Journal of Marketing Research*, XLII :233–239, May 2005.

[166] D.-I. Lin and Z. Kedem. Pincer-Search : An Efficient Algorithm for Discovering the Maximum Frequent Set. *IEEE Transactions on Knowledge and Data Engineering*, 14 :553–566, 2002.

[167] T. Y. Lin, Y. Y. Yao, and E. Louie. Value Added Association Rules. In *PAKDD '02 : Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 328–333, London, UK, 2002. Springer-Verlag.

[168] W. Lin, S. A. Alvarez, and C. Ruiz. Efficient Adaptive-Support Association Rule Mining for Recommender Systems. *Data Mining and Know. Disc.*, 6(1) :83–105, January 2002.

[169] G. Linden, B. Smith, and J. York. Amazon.com Recommendations : Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7 :76–80, January 2003.

[170] C. X. Ling, T. Chen, Q. Yang, and J. Cheng. Mining Optimal Actions for Profitable CRM. In *Proceedings of 2002 IEEE International Conference on Data Mining*, 2002.

[171] C. X. Ling and C. Li. Data mining for Direct Marketing : Problems and Solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 73–79. AAAI Press, 1998.

[172] R. Ling and D. C. Yen. Customer Relationship Management : An Analysis Framework and Implementation Strategies. *The Journal of Comp. Inf. Syst.*, 41 :82–97, 2001.

[173] B. Liu, W. Hsu, and S. Chen. Using General Impressions to Analyze Discovered Classification Rules. In *KDD*, pages 31–36, 1997.

[174] B. Liu, W. Hsu, and Y. Ma. Pruning and Summarizing the Discovered Associations. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD-99)*, pages 125–134. ACM Press, 1999.

[175] B. Liu, W. Hsu, and Y. Ma. Identifying Non-Actionable Association Rules. In *KDD '01 : Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–334, New York, NY, USA, 2001. ACM.

[176] D.-R. Liu, C.-H. Lai, and W.-J. Lee. A Hybrid of Sequential Rules and Collaborative Filtering for Product Recommendation. *Inf. Sciences*, 179(20) :3505–3519, September 2009.

[177] Y. Liu and M. Schumann. Data Mining Feature Selection for Credit Scoring Models. *Journal of the Operational Research Society*, 56 :1099–1108(10), September 2005.

[178] M. Lobur, Y. Stekh, A. Kernytskyy, and F. Sardieh. Some Trends in Knowledge Discovery and Data Mining. In *Perspective Technologies and Methods in MEMS Design 2008*, number 95-97, May 2008.

[179] O. Maimon and L. Rokach. *Data Mining and Knowledge Discovery Handbook*. Computer Science, 2005.

[180] Y. Malhotra. Knowledge Management and New Organization Forms : A Framework for Business Model Innovation. *Inf. Resour. Manage. J.*, 13 :5–14, January 2000.

[181] H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient Algorithms for Discovering Association Rules. In U. M. Fayyad and R. Uthurusamy, editors, *AAAI Workshop on Knowledge Discovery in Databases (KDD-94)*, pages 181–192, Seattle, Washington, 1994. AAAI Press.

[182] C. Marinica and F. Guillet. Knowledge-Based Interactive Postmining of Association Rules Using Ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 99(RapidPosts), 2010.

[183] B. Masand and G. Shapiro. A Comparison of Approaches for Maximizing Business Payoff of Prediction Models. In *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining*, pages 195–201. ACM, 1996.

[184] C. J. Matheus, G. Piatetsky-Shapiro, and D. Mcneill. An Application of KEFIR to the Analysis of Healthcare Information. In *In Proc. of the Eleventh International Conference on Artificial Intelligence, Workshop on Knowledge Discovery in Databases*, pages 25–36, 1994.

[185] C. J. Matheus, G. Piatetsky-Shapiro, and D. McNeill. Selecting and Reporting What Is Interesting. In *Advances in Knowledge Discovery and Data Mining*, pages 495–515. 1996.

[186] D. McSherry. Explanation in Recommender Systems. *Artif. Intell. Rev.*, 24(2) :179–197, 2005.

[187] R. B. Messaoud, O. Boussaid, and S. L. Rabaséda. Evaluation of a MCA-based Approach to Organize Data Cubes. In *CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 341–342, New York, NY, USA, 2005. ACM.

[188] R. V. Meteren and M. V. Someren. *Using Content-Based Filtering for Recommendation*. 2000.

[189] L. S. Meyers, G. Gamst, and A. J. Guarino. *Applied Multivariate Research : Design and Interpretation*. SAGE Publications, 2006.

[190] Micropole-Univers. Les Entreprises Satisfaites de leur Application de CRM. http ://www.itchannel.info/articles/95244/selon-etude-micropole-univers-br-entreprises-satisfaites-application-crm.html, Septembre 2009.

[191] S. E. Middleton, H. Alani, and D. C. D. Roure. Exploiting Synergy Between Ontologies and Recommender Systems. In *Proceedings of the WWW2002 International Workshop on the Semantic Web*, 2002.

[192] S. E. Middleton, N. R. Shadbolt, and D. C. D. Roure. Ontological User Profiling in Recommender Systems. *ACM Transactions on Information Systems*, 22 :54–88, 2004.

[193] B. N. Miller, I. Albert, S. K. Lam, J. A. Konstan, and J. Riedl. MovieLens Unplugged : Experiences With an Occasionally Connected Recommender System. In *Proceedings of the 8th international conference on Intelligent user interfaces*, IUI '03, pages 263–266, New York, NY, USA, 2003. ACM.

[194] B. N. Miller, J. A. Konstan, and J. Riedl. PocketLens : Toward a Personal Recommender System. *ACM Trans. Inf. Syst.*, 22 :437–476, July 2004.

[195] S. Mitra, S. K. Pal, and P. Mitra. Data Mining in Soft Computing Framework : A Survey. *IEEE Transactions on Neural Networks*, 13 :3–14, 2001.

[196] K. Miyahara and M. J. Pazzani. Collaborative Filtering with the Simple Bayesian Classifier. In *Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence*, pages 679–689, 2000.

[197] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective Personalization Based on Association Rule Discovery from Web Usage Data. In *Proceedings of the 3rd international workshop on Web information and data management*, WIDM '01, pages 9–15, New York, NY, USA, 2001. ACM.

[198] M. Montaner, B. López, and J. L. De La Rosa. A Taxonomy of Recommender Agents on the Internet. *Artif. Intell. Rev.*, 19 :285–330, June 2003.

[199] R. J. Mooney and L. Roy. Content-Based Book Recommending Using Learning for Text Categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, DL '00, pages 195–204, New York, NY, USA, 2000. ACM.

[200] R. Mourad, C. Sinoquet, and P. Leray. Learning Hierarchical Bayesian Networks for Genome-Wide Association Studies. In *19th International Conference on Computational Statistics (COMPSTAT)*, 2010.

[201] K. Nageswara Rao and V. G. Talwar. Application Domain and Functional Classification of Recommender Systems a Survey. *Desidoc journal of library and information technology*, 28(3) :17–36, 2008.

[202] P. Naïm, P.-H. Wuillemin, P. Leray, O. Pourret, and A. Becker. *Réseaux Bayésiens*. Eyrolles, Paris, 2004.

[203] M. Nakagawa and B. Mobasher. Impact of Site Characteristics on Recommendation Models Based On Association Rules and Sequential Patterns. *In Intelligent Techniques for Web Personalization*, 2003.

[204] E. Ngai. Customer Relationship Management Research (1992-2002) : An Academic Literature Review and Classification. *Marketing Intelligence and Planning*, 23(6) :582–605, 2005.

[205] E. Ngai, L. Xiu, and D. Chau. Application of Data Mining Techniques in Customer Relationship Management : A Literature Review and Classification. *Expert Systems with Applications*, 36(2) :2592–2602, March 2009.

[206] R. Niraj, M. Gupta, and C. Narasimhan. Customer Profitability in a Supply Chain. *Journal of Marketing Research*, 65 :1–16, 2001.

[207] E. R. Omiecinski. Alternative Interest Measures for Mining Associations in Databases. *IEEE Trans. on Knowl. and Data Eng.*, 15(1) :57–69, 2003.

[208] K. Onuma, H. Tong, and C. Faloutsos. TANGENT : a Novel, "Surprise me", Recommendation Algorithm. In *KDD '09 : Proc. of the 15th ACM SIGKDD Int. Conf. on KDD*, pages 657–666, New York, NY, USA, 2009. ACM.

[209] B. Padmanabhan and A. Tuzhilin. A Belief-Driven Method for Discovering Unexpected Patterns. In *KDD*, pages 94–100, 1998.

[210] J. Pagès. Analyse Factorielle de Données Mixtes. *Revue Statistique Appliquée*, LII :93–111, 2004.

[211] Y.-J. Park and A. Tuzhilin. The Long Tail of Recommender Systems and How to Leverage It. In *Proceedings of the 2008 ACM conference on Recommender Systems*, RecSys '08, pages 11–18, New York, NY, USA, 2008. ACM.

[212] K. Parsaye. The Sandwich Paradigm. *Database Programming & Design*, pages 50–55, April 1995.

[213] A. Parvatiyar and J. N. Sheth. Customer Relationship Management : Emerging Practice, Process and Discipline. *Journal of Economic and Social Research*, 3 :6–23, 2002.

[214] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient Mining of Association Rules Using Closed Itemset Lattices. *Inf. Syst.*, 24(1) :25–46, 1999.

[215] M. J. Pazzani. A Framework for Collaborative, Content-Based and Demographic Filtering. *Artif. Intell. Rev.*, 13 :393–408, December 1999.

[216] M. J. Pazzani and D. Billsus. The Adaptive Web. chapter Content-Based Recommendation Systems, pages 325–341. Springer-Verlag, Berlin, Heidelberg, 2007.

[217] P. Pendharkar. *Managing Data Mining Technologies in Organizations : Techniques and Applications*. Idea Group Publishing, 2002.

[218] D. M. Pennock, E. Horvitz, and C. L. Giles. Social Choice Theory and Recommender Systems : Analysis of The Axiomatic Foundations of Collaborative Filtering. In *Proc. 17th AAAI*, 2000.

[219] G. Piatetsky-Shapiro. *Discovery, Analysis, and Presentation of Strong Rules*. AAAI/MIT Press, Cambridge, MA, 1991.

[220] G. Piatetsky-Shapiro. Data Mining and Knowledge Discovery 1996 to 2005 : Overcoming the Hype and Moving from University to Business and Analytics. *Data Min. Knowl. Discov.*, 15(1) :99–105, 2007.

[221] G. Piatetsky-Shapiro and C. J. Matheus. The Interestingness of Deviations. In *Proceedings of KDD-94 workshop*. AAAI Press, 1994.

[222] I. Pilászy and D. Tikk. Recommending New Movies : Even a Few Ratings Are More Valuable Than Metadata. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 93–100, New York, NY, USA, 2009. ACM.

[223] T. Piton, J. Blanchard, H. Briand, and F. Guillet. Domain Driven Data Mining to Improve Promotional Campaign ROI and Select Marketing Channels. In *The 18th ACM Conference on Information and Knowledge Management The 18th ACM Conference on Information and Knowledge Management*, pages 1057–1066, Hong-Kong, 2009. ACM.

[224] T. Piton, J. Blanchard, H. Briand, L. Tessier, and G. Blain. Analyse et Application de Modèles de Régression Pour Optimiser le Retour sur Investissement d'Opérations Commerciales. In *Actes des neuvièmes journées Extraction et Gestion des Connaissances EGC'2009 Extraction et gestion des connaissances (EGC'2009)*, Revue des Nouvelles Technologies de l'Information, pages 25–30, Strasbourg France, 2009.

[225] T. Piton, J. Blanchard, and F. Guillet. CAPRE : A New Methodology for Product Recommendation Based on Customer Actionability and Profitability. In *In proceedings of the fifth International Workshop on Domain Driven Data Mining (DDDM 2011) in conjunction with IEEE ICDM 2011*, 2011.

[226] T. Piton, J. Blanchard, F. Guillet, and H. Briand. Une méthodologie de recommandations produits fondée sur l'actionnabilité et l'intérêt économique des clients. In Ali Khenchaf et Pascal Poncelet, editor, *Actes des onzièmes journées Extraction et Gestion des Connaissances EGC'2011 Extraction et gestion des connaissances (EGC'2011)*, volume E-20 of *Revue des Nouvelles Technologies de l'Information*, pages 203–214, Brest France, 01 2011. Hermann.

[227] D. Poirier. *Des Textes Communautaires à la Recommandation*. Thèse, en convention CIFRE entre l'entreprise Orange, le laboratoire d'Informatique Fondamentale d'Orléans et le Laboratoire d'Informatique de Paris 6, Orléans, France, Février 2011. Isabelle Tellier et Patrick Gallinari (Dir).

[228] P. Pu and L. Chen. Trust Building with Explanation Interfaces. In *IUI '06 : Proceedings of the 11th international conference on Intelligent user interfaces*, pages 93–100, New York, NY, USA, 2006. ACM.

[229] B. Rahul and Z. Yi. CRM Systems Used for Targeting Market : A Case at Cisco Systems. In *ICEBE '05 : Proceedings of the IEEE International Conference on e-Business Engineering*, pages 183–186, Washington, DC, USA, 2005. IEEE Computer Society.

[230] Z. W. Ras and A. Dardzinska. Action Rules Discovery, a New Simplified Strategy. In *Foundations of Intelligent Systems*, pages 445–453. Springer, 2006.

[231] Z. W. Ras and A. Wieczorkowska. Action-Rules : How to Increase Profit of a Company. In *PKDD*, pages 587–592, 2000.

[232] Z. W. Ras, E. Wyrzykowska, and H. Wasyluk. ARAS : Action Rules Discovery Based on Agglomerative Strategy. In *MCD*, pages 196–208, 2007.

[233] F. F. Reichheld and W. E. Sasser. Zero Defections : Quality Comes to Services. *Harvard Business Review*, Sept-Oct 1990.

[234] W. Reinartz, J. S. Thomas, and V. Kumar. Balancing Acquisition and Retention Resources to Maximize Customer Profitability. *Journal of Marketing*, 69(1) :63–79, January 2005.

[235] M. Ren, Z. Chen, C. Liu, and G. Chen. An Evolving Information System Based on Data Mining Knowledge to Support Customer Relationship Management. In *Advanced Management of Information for Globalized Enterprises*, pages 1–5, September 2008.

[236] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens : an Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the*

1994 ACM conference on Computer supported cooperative work, CSCW '94, pages 175–186, New York, NY, USA, 1994. ACM.

[237] P. Resnick and H. R. Varian. Recommender Systems. *Communications of the ACM*, 40(3) :56–58, 1997.

[238] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[239] E. Rich. Readings in Intelligent User Interfaces. chapter User Modeling via Stereotypes, pages 329–342. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[240] J. F. Roddick and S. Rice. What's Interesting About Cricket ? On Thresholds and Anticipation in Discovered Rules. *SIGKDD Explorations*, 3(1) :1–5, 2001.

[241] L. Ryals. Creating Profitable Customers Through the Magic of Data Mining. *Journal of Targeting, Measurement and Analysis for Marketing*, 11 :343–349(7), May 2003.

[242] C. Rygielski, J.-C. Wang, and D. C. Yen. Data Mining Techniques for Customer Relationship Management. *Technology in Society*, 24(4) :483 – 502, 2002.

[243] S. Sahar. Interestingness via What is Not Interesting. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 332–336, New York, NY, USA, 1999. ACM.

[244] G. Saporta. Simultaneous Analysis of Qualitative and Quantitative Data. In *Proceedings of the 35th Scientific Meeting of the Italian Statistical Society*, pages 63–72, 1990.

[245] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-Based Collaborative Filtering Recommendation Algorithms. In *WWW '01 : Proceedings of the 10th international conference on World Wide Web*, pages 285–295, New York, NY, USA, 2001. ACM.

[246] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of Recommendation Algorithms for E-Commerce. In *EC '00 : Proceedings of the 2nd ACM conference on Electronic commerce*, pages 158–167, New York, NY, USA, 2000. ACM.

[247] A. Savasere, E. Omiecinski, and S. Navathe. An Efficient Algorithm for Mining Association Rules in Large Databases. In *Proceedings of the 21st VLDB Conference*, pages 432–443, Zurich, Switzerland, 1995.

[248] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. The Adaptive Web. chapter Collaborative Filtering Recommender Systems, pages 291–324. Springer-Verlag, Berlin, Heidelberg, 2007.

[249] J. B. Schafer, A. J. Konstan, and J. Riedl. E-Commerce Recommendation Applications. *Data Mining and Knowledge Discovery*, 5(1 - 2) :115–153, January 2001.

[250] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Generative Models for Cold-Start Recommendations. In *Proceedings of the 2001 SIGIR Workshop on Recommender Systems*, 2001.

[251] C. Schmitt. Vers l'Actionnabilité de la Recherche en Gestion : pour une Approche Dialectique entre Recherche et Pratiques. *Academy of Management*, Division Méthodes de Recherche, 2004.

[252] A. Scouarnec. L'Observation des Métiers : Définition, Méthodologie et Actionnabilité en GRH. *Management et Avenir*, 1 :23–42, 2004.

[253] G. Shani, M. Chickering, and C. Meek. Mining Recommendations from the Web. In *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys, pages 35–42, New York, NY, USA, 2008. ACM.

[254] U. Shardanand and P. Maes. Social Information Filtering : Algorithms for Automating Word of Mouth. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '95, pages 210–217, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[255] M. J. Shaw, C. Subramaniam, G. W. Tan, and M. E. Welge. Knowledge Management and Data Mining for Marketing. *Decision Support Systems*, 31(1) :127 – 137, 2001.

[256] C. Shearer. The CRISP-DM Model : The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4), 2000.

[257] B. Shekar and R. Natarajan. A Fuzzy Relatedness Measure for Determining Interestingness of Association Rules. In *HIS*, pages 95–104, 2002.

[258] Y.-D. Shen, Z. Zhang, and Q. Yang. Objective-Oriented Utility-Based Association Mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 426–433, 2002.

[259] A. Silberschatz and A. Tuzhilin. On Subjective Measures of Interestingness in Knowledge Discovery. In *KDD*, pages 275–281, 1995.

[260] A. Silberschatz and A. Tuzhilin. What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Transactions on Knowledge and Data Engineering*, 8 :970–974, 1996.

[261] S. J. Simoff. Towards the Development of Environments for Designing Visualisation Support for Visual Data Mining. In *Proceedings Int. Workshop on Visual Data Mining, 12th European Conference on Machine Learning and 5th European Conference on Principles and Practice of Knowledge Discovery in Databases ECML/PKDD2001*, pages 93–106, 2001.

[262] B. Smyth, K. McCarthy, J. Reilly, D. O'Sullivan, L. McGinty, and D. C. Wilson. Case Studies in Association Rule Mining for Recommender Systems. In *IC-AI*, pages 809–815, 2005.

[263] M. Spiliopoulou and C. Pohle. Data mining for Measuring and Improving the Success of Web Sites. *Data Mining and Knowledge Discovery*, 5(1-2) :85–114, JAN-APR 2001.

[264] R. Srikant and R. Agrawal. Mining Generalized Association Rules. In *Proceedings of the 21st VLDB Conference, Zurich, Switzerland*, 1995.

[265] X. Su and T. M. Khoshgoftaar. A Survey of Collaborative Filtering Techniques. *Adv. in Artif. Intell.*, page 42, January 2009.

[266] K. Swearingen and R. Sinha. Interaction Design for Recommender Systems. In *In Designing Interactive Systems 2002. ACM*. Press, 2002.

[267] R. Swift. *Accelerating Customer Relationships : Using CRM and Relationship Technologies*. Prentice Hall Press, Upper Saddle River, NJ, USA, 2001.

[268] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the Right Objective Measure for Association Analysis. *Information Systems*, 29(4) :293–313, 2004.

[269] T. S. H. Teo, P. R. Devadoss, and S. L. Pan. Towards a Holistic Perspective of Customer Relationship Management (CRM) Implementation : A Case Study of the Housing and Development Board. *Decision Support Systems*, 42(3) :1613–1627, 2006.

[270] C. A. Thompson, M. H. Goker, and P. Langley. A Personalized System for Conversational Recommendations. *Journal of Artificial Intelligence Research*, 21 :393–428, 2004.

[271] O. Thonnard and M. Dacier. Actionable Knowledge Discovery for Threats Intelligence Support Using a Multi-dimensional Data Mining Methodology. In *Proceedings of the 2008 IEEE International Conference on Data Mining Workshops*, pages 154–163, Washington, DC, USA, 2008. IEEE Computer Society.

[272] N. Tintarev and J. Masthoff. A Survey of Explanations in Recommender Systems. *Data Engineering Workshops, 22nd International Conference on*, 0 :801–810, 2007.

[273] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila. Pruning and Grouping Discovered Association Rules. In *ECML'95 MLnet workshop on statistics, machine learning, and knowledge discovery in databases*, pages 47–52, April 1995.

[274] R. T. Trevor Hastie and J. Friedman. *The Elements of Statistical Learning : Data Mining, Inference and Prediction*. Springer, 2 edition, February 2009.

[275] S. Tufféry. *Data Mining et Statistique Décisionnelle, l'Intelligence dans les Bases de Données*. 3 edition, 2010.

[276] S. Tufféry. *Étude de Cas en Statistique Décisionnelles*. 3 edition, 2010.

[277] E. Turban, J. E. Aronson, T.-P. Liang, and R. Sharda. *Decision Support and Business Intelligence Systems*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 8 edition, 2006.

[278] A. Tuzhilin and G. Adomavicius. Discovery of Actionable Patterns in Databases : The Action Hierarchy Approach. In *KDD*, pages 111–114. AAAI Press, 1997.

[279] A. A. Tzacheva and Z. W. Raś. Action Rules Mining : Research Articles. *Int. J. Intell. Syst.*, 20(7) :719–736, 2005.

[280] P. Tzvetkov, X. Yan, and J. Han. TSP : Mining Top-K Closed Sequential Patterns. *Data Mining, IEEE International Conference on*, 0 :347, 2003.

[281] L. Ungar and D. Foster. Clustering Methods For Collaborative Filtering. In *Proceedings of the Workshop on Recommendation Systems*. AAAI Press, Menlo Park California, 1998.

[282] B. Vaillant, P. Lenca, and S. Lallich. A Clustering of Interestingness Measures. In *Discovery Science*, pages 290–297, 2004.

[283] S. Vanderlinden. Customer is King. *Claims Magazine*, September 2009.

[284] K. Vanhoof, P. Pauwels, J. Dombi, T. Brijs, and G. Wets. Penalty-Reward Analysis with Uninorms : A Study of Customer (Dis)Satisfaction. In *Intelligent Data Mining*, pages 237–252. 2005.

[285] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[286] K. V. Venkatesan, R. A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. *Journal of Mark. Research*, 68 :106–125, October 2004.

[287] Z. Wan. Personalized Tourism Information System in Mobile Commerce. *Management of e-Commerce and e-Government*, pages 387–391, 2009.

[288] J. Wang, J. Han, Y. Lu, and P. Tzvetkov. TFP : An Efficient Algorithm for Mining Top-K Frequent Closed Itemsets. *IEEE Transactions on Knowledge and Data Engineering*, 17(5) :652–664, 2005.

[289] K. Wang, Y. Jiang, and A. Tuzhilin. Mining Actionable Patterns by Role Models. *Data Engineering 2006 ICDE 06 Proceedings of the 22nd International Conference on*, page 16, 2006.

[290] K. Wang, S. Zhou, and J. Han. Profit Mining : From Patterns to Actions. In *Proceedings of the 8th International Conference on Extending Database Technology : Advances in Database Technology*, EDBT '02, pages 70–87, London, UK, 2002. Springer-Verlag.

[291] K. Wang, S. Zhou, and G. Webb. Mining Customer Value : from Association Rules to Direct Marketing. *In Proceedings of the IEEE International Conference on Data Engineering*, 11 :57–80, 2005.

[292] Y. Wang, Z. Li, and Y. Zhang. Mining Sequential Association-Rule for Improving WEB Document Prediction. In *Proceedings of the Sixth International Conference on Computational Intelligence and Multimedia Applications*, pages 146–151, Washington, DC, USA, 2005. IEEE Computer Society.

[293] G. I. Webb. Discovering Significant Rules. In *KDD '06 : Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 434–443, New York, NY, USA, 2006. ACM.

[294] C.-P. Wei and I.-T. Chiu. Turning Telecommunications Call Details to Churn Prediction : a Data Mining Approach. *Expert Systems with Applications*, 23(2) :103 – 112, 2002.

[295] C. Westphal and T. Blaxton. *Data Mining Solutions : Methods and Tools for Solving Real-World Problems*. John Wiley & Sons, 1998.

[296] K. Wickramaratna, M. Kubat, and K. Premaratne. Predicting Missing Items in Shopping Carts. *IEEE Trans. on Knowl. and Data Eng.*, 21 :985–998, July 2009.

[297] N. Wingfield and J. Pereira. Amazon Uses Faux Suggestions to Promote New Clothing Store. *Wall Street Journal (December 4)*, December 2002.

[298] D. Xin, H. Cheng, X. Yan, and J. Han. Extracting Redundancy-Aware Top-k Patterns. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 444–453, New York, NY, USA, 2006. ACM.

[299] S. B. Yahia and E. M. Nguifo. Approches d'Extraction de Règles d'Association Basées sur la Correspondance de Galois. *Ingénierie des Systèmes d'Information*, 9(3-4) :23–55, 2004.

[300] L. Yan and P. Baldasare. Beyond Classification and Ranking : Constrained Optimization of the ROI. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 948–953, New York, NY, USA, 2006. ACM.

[301] Q. Yang, J. Yin, C. Ling, and T. Chen. Postprocessing Decision Trees to Extract Actionable Knowledge. In *Third IEEE International Conference on Data Mining*, pages 685–688. IEEE Computer Soc, 2003.

[302] Q. Yang, J. Yin, C. Ling, and R. Pan. Extracting Actionable Knowledge from Decision Trees. *IEEE Transactions On Knowledge and Data Engineering*, 19(1) :43–56, Jan 2007.

[303] Y. Yinghui. *New Data Mining and Marketing Approaches for Customer Segmentation and Promotion Planning on the Internet*. PhD thesis, Philadelphia, PA, USA, 2004.

[304] M. J. Zaki and M. Ogihara. Theoretical Foundations of Association Rules. pages 1–8, June 1998.

[305] M. J. Zaki. Generating Non-Redundant Association Rules. *International Conference on Knowledge Discovery and Data Mining*, pages 34 – 43, 2000.

[306] M. J. Zaki and C.-J. Hsiao. CHARM : An Efficient Algorithm for Closed Itemset Mining. In *Proceedings of the Second SIAM DM*, Arlington, VA, 2002. SIAM.

[307] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New Algorithms for Fast Discovery of Association Rules. In *KDD*, pages 283–286, 1997.

[308] M. Zeller, R. Grossman, C. Lingenfelder, M. R. Berthold, E. Marcade, R. Pechter, M. Hoskins, W. Thompson, and R. Holada. Open Standards and Cloud Computing : KDD-2009 Panel Report. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 11–18, New York, NY, USA, 2009. ACM.

[309] Y. Zhang and J. R. Jiao. An Associative Classification-Based Recommendation System for Personalization in B2C e-commerce Applications. *Expert Syst. Appl.*, 33(2) :357–367, 2007.

[310] K. Zhao, B. Liu, T. M. Tirpak, and W. Xiao. Opportunity Map : a Visualization Framework for Fast Identification of Actionable Knowledge. In *CIKM*, pages 60–67, 2005.

[311] Y. Zhao, C. Zhang, and L. Cao. *Post-Mining of Association Rules : Techniques for Effective Knowledge Extraction*. IGI Global, 2009.

[312] Y. Zhao, H. Zhang, L. Cao, C. Zhang, and H. Bohlscheid. Combined Pattern Mining : From Learned Rules to Actionable Knowledge. In *AI 08 : Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence*, pages 393–403, Berlin, Heidelberg, 2008. Springer-Verlag.

[313] H. Zhu, P. A. Beling, and G. A. Overstreet. A Study in the Combination of two Consumer Credit Scores. *Journal of the Operational Research Society*, 52(9) :974 – 980, 2001.

[314] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving Recommendation Lists Through Topic Diversification. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA, 2005. ACM.

Résumé : Dans un contexte concurrentiel, la richesse des entreprises réside dans leurs clients. Il est plus rentable de fidéliser un client existant que d'en acquérir un nouveau. De ce fait, les entreprises cherchent à mieux connaître leurs clients pour trouver des moyens de les fidéliser. Cette approche de la connaissance des clients fondée sur l'analyse des données se heurte toutefois au volume important des données. Ce constat pousse les entreprises à Extraire des Connaissances à partir des Données. Ces connaissances et leur actionnabilité fournissent aux experts un outil d'aide à la décision dont la performance peut être mesurée par le retour sur investissement généré par les actions. Les systèmes de recommandation sont adaptés pour mettre en place ces outils car ils permettent de filtrer l'information puis de recommander de manière proactive des produits susceptibles de fidéliser le client. Dans le cadre d'une stratégie commerciale basée sur les forces de vente, comment fidéliser les clients pour accroître leur valeur ? Une mauvaise recommandation intrusive peut en effet avoir des répercussions importantes sur le client et le commercial peut refuser d'utiliser le système s'il ne juge pas les recommandations suffisamment pertinentes. Pour s'affranchir de ces contraintes, nous avons proposé la méthodologie *CAPRE* qui consiste à extraire des comportements de référence sous la forme de cohortes de règles en ciblant raisonnablement les clients présentant un manque à gagner et en quantifiant le profit espéré. Cette approche a été mise en œuvre au sein de l'outil *ARKIS*. Notre méthodologie a été validée sur le jeu de données *MovieLens* puis validée et appliquée sur les données opérationnelles du groupe VM Matériaux.

Mots-clés :
Extraction de connaissances à partir des données, *gestion de la relation client*, *actionnabilité*, *profitabilité*, *système de recommandation*, *fouille de données pour le marketing*, *application industrielle*.

Abstract :
In an extremely competitive market, companies' wealth is their customers. It is more profitable to retain existing customers than to acquire new ones. Therefore, companies seek ways to retain their customers by understanding them better. This approach based on data analysis faces the large volumes of data and leads companies to extract knowledge from data. Both knowledge and actionability provide a decision-making tool for experts whose performance can be measured by the Return On Investment generated by actions. Recommender systems are designed to implement these tools as they allow companies to filter information and recommend products to customers according to their preferences. As part of a business strategy based on the sales force, how can customers be retained and their value increased ? The cost of an inappropriate recommendation is higher for salespersons' visits than for e-commerce websites. Salespeople may even refuse to use the system if they find the recommendations not sufficiently relevant. To overcome these limitations, we propose *CAPRE*, a new methodology for recommender systems based on the analysis of turnover for customers of specific products. CAPRE aggregates rules to extract characteristic purchasing behaviors, and then analyzes the counter-examples to detect the most actionable and profitable customers. Recommendations are made by targeting the actionable counter-examples with the most profitable rules. This approach has been implemented in the *ARKIS* tool. We measured the effectiveness of our recommender system on the MovieLens benchmark with a cross validation and applied it to over 10,000 customers and 100,000 products of *VM Matériaux* company.

Keywords :
Actionable knowledge discovery, *customer relationship management*, *actionability*, *profitability*, *recommender system*, *data mining for marketing*, *industrial case-based application*.